Computer Science Review 58 (2025) 100782

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

The computing continuum: Past, present, and future

Luiz F. Bittencourt [a,*], Roberto Rodrigues-Filho [b], Josef Spillner [c], Filip De Turck [d], José Santos [d], Nelson L.S. da Fonseca [a], Omer Rana [e], Manish Parashar [f], Ian Foster [g]

[a] *Instituto de Computação, Universidade Estadual de Campinas (UNICAMP), Brazil*
[b] *Departamento de Computação, Universidade Federal de Santa Catarina (UFSC), Brazil*
[c] *School of Engineering, Zurich University of Applied Sciences, Switzerland*
[d] *IDLab, Ghent University, Belgium*
[e] *School of Computer Science and Informatics, Cardiff University, UK*
[f] *Scientific Computing and Imaging (SCI) Institute, University of Utah, USA*
[g] *University of Chicago & Argonne National Laboratory, USA*

ARTICLE INFO

Keywords:
Computing continuum
Cloud computing
Edge computing
Fog computing
Internet of Things

ABSTRACT

The development of network-connected computing resources has led to various computing paradigms over the years, each bringing its own set of challenges for creating efficient distributed systems. Currently, there is an increasing need to integrate the evolving Internet of Things (IoT) with the established Cloud infrastructure. This integration often requires adding intermediate layers to address Cloud limitations such as latency, bandwidth, security, cost, and control. This configuration, known as the computing continuum, involves a diverse array of distributed devices with unique characteristics working together to meet the demands of both current and emerging applications. This paper explores the path that has led to the development of the computing continuum, offering a technology-agnostic definition from a historical perspective. It also examines applications that can benefit from the computing continuum and identifies research challenges that need to be addressed to fully realize its potential.

Contents

[*] Corresponding author.
E-mail addresses: bit@ic.unicamp.br, bit@unicamp.br (L.F. Bittencourt).

https://doi.org/10.1016/j.cosrev.2025.100782
Received 27 January 2025; Received in revised form 28 May 2025; Accepted 14 June 2025
Available online 10 July 2025

1. Introduction

Information technology systems were born as centralized entities, consisting of *mainframes* computers, that then became embedded in, and were often replaced by, distributed computing infrastructures made possible by the development of computing networks. These developments allowed the Internet to connect widely distributed computing resources to serve users worldwide. A few years ago, a server on the Internet typically worked on a request–response pattern, transmitting a low amount of information per request due to network capacity constraints. As communications technologies continued to evolve, the capacity for data traffic on the Internet increased many times, allowing servers to receive and transmit large amounts of data and enabling remote execution of a plethora of data-demanding applications. These developments culminated in the cloud computing paradigm, in which a relatively small number of data centers are dedicated to data storage and hosting and/or processing applications on demand [1]. Cloud computing offers remote computing capacity that can be used by any electronic device connected to the Internet, large or small, transparently to the user.

Concomitantly with the evolution of computer networks, electronic devices have been expanding their ability to generate data, resulting in the accumulation of a tremendous variety of information, from measurements of natural phenomena to data generated as a result of human actions. The Internet of Things (IoT) [2] anticipates an unprecedented number of devices scattered at the edges of the network that are expected to connect with the broader Internet. Along with this connection of *everything* to the Internet comes the need to transfer, store, and process unprecedented amounts of data — a need that lays the path for research in distributed computing infrastructures. Cloud computing is now consolidated as a paradigm to fulfill the requirements of many applications, from IoT to scientific computing.

IoT devices rely on cloud computing for data management, information, and knowledge production. Many applications experience a fast time to market with the widespread adoption of cloud computing, taking advantage of the flexibility of the paradigm and reduced initial capital expenditure. However, the limitations of the cloud computing paradigm were exposed after its wide adoption, as centralized servers cannot always fulfill all the requirements of several classes of modern and future applications, such as real-time, interactive, low latency, bandwidth-intensive, and mobile [3]. Some limitations are inherent in centralization: the physical distance between devices and users introduces delays, further increased by the delay due to communication that traverses multiple hops (and routers) from the edge to the cloud [4]. In the face of this, resource selection for data distribution and processing should ideally consider both networking and computing infrastructures.

With large data sets being generated and consumed at the network's edge and with the widespread adoption of cloud computing, edge devices are now combined with the cloud to run heterogeneous applications and consume data from various sources. Combining the cloud with the ever-increasing ability of edge devices to process data requires novel distributed computing infrastructures that can cope with such heterogeneous application requirements [3]. To overcome the limitations of centralization, edge computing has been proposed to bring computing capacity closer to the edge of the network [5], improving aspects such as response time and reducing aggregated bandwidth use. Combining the ability to run smaller, localized applications at the edge with high cloud capacity, fog computing has emerged as a paradigm that can support heterogeneous requirements of small and large applications through multiple layers of this computing infrastructure [6,7]. Today, edge and fog computing employ IoT devices to reduce limitations of the cloud paradigm by using scattered resources to capture and process data.

As computational and data resources within a distributed computing system become yet more scattered, networking and connectivity among devices becomes more relevant to the performance of applications. The evolution of the current global IT infrastructure points toward a widely distributed network of computing capacity connected by ultra-fast networks. As the connection speed among devices approximates that of parallel computing, the *computing continuum* emerges. Moreover, latencies bound by laws of physics play a fundamental role in distinguishing local and distributed computing resources in the *continuum*. On the other hand, bandwidth importance is reduced in the resource allocation decision-making process for many applications: it often becomes more important to select the more efficient (e.g., special-purpose) resource for an application, where the gain in processing time is greater than the time taken to transfer data and results. This scenario dramatically increases the number of possible choices for allocating applications to resources that meet their demands, increasing the complexity (due to an increase in choice of potential candidate resources) of the resource allocation problem. At the same time, the possibility of *computing anywhere* brings more flexibility to resource allocation, making a variety of objective functions achievable (e.g., faster, greener, cheapest, or most secure) without necessarily hindering application performance or quality of service requirements.

Surveys and definitions of the computing continuum in existing literature are often limited to a study on the composition of IoT, edge/fog, and cloud paradigms, which is reinforced by the broad use of the *Cloud continuum* terminology. In this paper, we visit the historical perspective of distributed computing leading towards the computing continuum, review the computing continuum literature, and provide a technology-agnostic definition of the continuum, encompassing the joint use of networking and computing devices seamlessly to offer a pervasive computing infrastructure. We argue that such a composition of resources is more complex to manage than a composition of existing distributed computing paradigms, demanding efforts beyond resource management but also integrating computer networking and

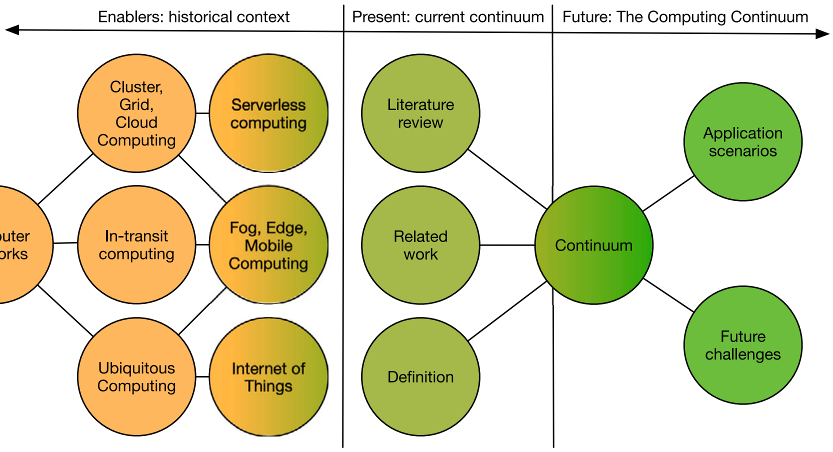

Fig. 1. Topics covered and organization of this paper.

programming aspects. We also discuss application use cases and future challenges in making the computing continuum a reality for developers and users.

This paper is organized to present first, in Section 2, a historical perspective of enablers of the computing continuum and the present state of the art, with a literature review of computing continuum papers and a related work section of other literature reviews in this context. We present our Computing Continuum definition in Section 3, followed by the future continuum perspectives, with application scenarios for the computing continuum in Section 4 and a discussion in Section 5 on future challenges to explore the computing continuum in its full extent. Fig. 1 provides an illustration of this paper contents and organization.

2. Past and present

This section first places the continuum in a historical context, covering networks and distributed computing infrastructures from an evolutionary perspective. We discuss how the network impacted users' access to such systems. We also review recent literature, considering different forms of computing continuum and emphasizing how this paper differentiates from current paradigms encountered in the continuum literature.

2.1. Historical context

This section aims to define and discuss existing distributed computing infrastructures and to understand how several different characteristics of these infrastructures and their network components are transposed into the computing continuum. By capturing this historical context and identifying similarities, we aim to identify perspectives from different research communities towards a common terminology that will allow integrating efforts into a seamless computing continuum.

2.1.1. Computer networks evolution

Distributed computing extends multiprocess communication across hardware boundaries, effectively requiring computer networks as physical carriers upon which application-specific messages can be exchanged. Over time, network hardware has evolved from simply wired cables into diverse wired and wireless mediums, and networking software that started as simple client/server socket interfaces has become highly capable structures for software-defined topologies, stream processing, and complex event processing. This so-called *network softwarization* transformed the network from a neutral bit carrier into an adaptative distributed processing infrastructure dedicated to dynamically tailoring network capacity and performance to application needs. This characteristic of softwarized networking enables a non-trivial coalescing of computing and networking infrastructures towards the computing continuum, whose joint design, programmability, and management have been neglected so far by distributed computing paradigms, including current literature focusing on IoT, edge and cloud computing continuum integration. This discussion is explored further in the literature review section.

Network hardware determines the characteristics of the underlying protocol, including bandwidth and throughput, latency and jitter, reliability, cost, and energy efficiency. In the computing continuum, the machine-to-machine (M2M) communication spectrum typically starts with wireless sensor networks such as Narrow-Band IoT (NB-IoT), LoRa (WAN), Bluetooth/BLE, and extends to mobile phone communication (i.e., 5G and 6G) to fixed networks such as 10 to 400 Gbit/s Ethernet. Consequently, today's distributed applications typically translate between low-level messages and higher-level protocols, including 0MQ, AMQP, MQTT, CoAP, HTTP, and ICN approaches [8]. This translation adds latency; therefore, the use of network resources should be avoided by appropriate software logic as long as local processing capacity is available. Due to network and workload dynamics, this logic must be self-learning and thus possess certain forms of intelligent or autonomic behavior, which should be integrated into a continuum management framework encompassing both networking and computing infrastructures.

Such autonomic network technologies for smart continuums have been investigated in numerous research projects (e.g., SELFNET [9] and NEPHELE [10]). For example, in 2016, SELFNET started studying a self-organizing network management framework for 5G use cases by combining virtualization and software-defined networking technologies with Artificial Intelligence (AI) capabilities, aiming to enable automated network monitoring. Recently, the 2022–25 European project NEPHELE aims to "enable the efficient, reliable, and secure end-to-end orchestration of hyper-distributed applications over programmable infrastructures that span across the compute continuum from Cloud-to-Edge-to-IoT introducing automation and decentralized intelligence mechanisms powered by 5G and distributed AI technologies" [10]. This automated, intelligent management of networks is also embraced in the 6G conceptualization, where distributed machine learning frameworks are expected to act in the network configuration and adaptation [11]. We expect such intelligent management frameworks to seamlessly interoperate with machine learning techniques in computing infrastructure management to compose the management framework for the computing continuum.

2.1.2. Cluster computing

As computing resources and local area networking became more affordable, multiple computers were connected while running the same software environment to execute computational jobs. Computing clusters have emerged with combined computing and storage capacity that have increased with the number of computers, limited according to Amdahl's law. Each extra computer yields a smaller performance gain due to limited scalability in networking and job scheduling overheads. However, since each computer still adds additional capacity, large-scale computing clusters became necessary for distributing compute-intensive jobs. From an economic perspective, they required more setup and maintenance work but were still more affordable than the only possible alternative, integrated high-performance computers. Distributed operating system approaches were attempted (e.g., MOSIX [12]). They have reduced in importance due to the dominance of distributed middleware on multiple independent operating systems [13].

Cluster computing became a popular distributed processing infrastructure with the Beowulf cluster built in the mid-1990's and used to support processing-intensive applications [14]. Today, most cluster computing occurs through an additional abstraction layer on top of virtualized resources (e.g., virtual machines and containers) distributed over physical machines concealed in a single facility. In this setup, popular distributed computing frameworks, such as Apache Spark and Flink, can be used assuming an underlying static compute cluster with n workers, $n \times m$ CPU cores, and $n \times k$ amounts of main memory. This assumption may no longer be sufficient with the proliferation of more dynamic environments and more heterogeneous hardware and connectivity across the continuum, where the integration of cluster computing facilities into a continuum of resource management efforts along with other distributed computing and networking infrastructures is necessary to enable the seamless integration of the computing continuum.

2.1.3. Grid computing

Conventional compute clusters required physical setup in terms of installation and wiring, as well as software setup, primarily concurrent software updates and user management in homogeneous infrastructures. Users who need cluster resources were much more keen to define their jobs, data, and deadlines. Hence, institutions started offering hosted cluster services that could be federated across computers and beyond institutional boundaries: the Grid computing, which became popular in the late 1990's. Grid systems such as Globus Toolkit, Ourgrid, Grid 5000, and others have achieved not only fully managed job execution but also checkpointing, execution monitoring, and complex workflow integration, thus opening up capable heterogeneous compute systems to collaborative researchers from a broad variety of fields, based on a translational approach [15]. However, grid systems typically act in batch mode with heavy-weight semantics, including the lack of performance guarantees for bursts of small jobs that require execution. Many resource management approaches developed for grid computing incorporated heterogeneous systems complexities and interoperability efforts, advancing from more homogeneous setups to cluster computing. Knowledge built from grid computing research and development will undoubtedly help glue together highly heterogeneous computing and networking elements to support effortless data management and code mobility in the computing continuum.

2.1.4. Ubiquitous computing and the internet of things

Ubiquity refers to the ability to obtain computing resources and services anytime, considering embedded computing and mobility [16]. Thus, ubiquitous computing is closely related to utility computing, which adds an on-demand, pay-per-use dimension, and pervasive computing. Ubiquitous computing for humans is further supported by the prevalent spread of portable mobile devices, primarily smartphones, that are instantly available to decode QR codes or to sign up for a nearby event, indicated by a hidden Bluetooth Low Energy (BLE) beacon as part of a pervasive support infrastructure along with WiFi and other connectivity options.

The IoT concept [2] appeared in the early 2000's, becoming slowly and increasingly popular until its explosion in the mid 2010's. With massive deployments of small edge nodes driven by IoT, IIoT, and other trends, nearby access to computing resources has become feasible, especially in urban or in-house environments. Nevertheless, this still precludes access to ready-to-use services. Most data is collected on those nodes but offloaded to centralized compute facilities without smart decisions on when to retain which data and for how long. In the computing continuum, data processing should be available both in computing devices and the network, and IoT data retention and offloading decision-making should be transparent and dynamic.

2.1.5. Cloud computing

Cloud computing emerged in the mid 2000's and became widely popular around 2010, and is still broadly adopted nowadays. Cloud computing services are based on centralized data centers, where computing capacity is offered by computing cluster virtualization deployed in buildings specially designed to host them [3]. Datacenter hosts are often connected via Ethernet; different topologies for this interconnect are available in the literature [17]. The internal data center infrastructure details and network topology are usually not disclosed to cloud computing clients, who are not concerned about these details even though they can impact the application's behavior. Significant impacts from infrastructure management and control should be stated and reflected in service level agreements (SLAs) between providers and users.

Cloud data centers are typically extensive facilities deployed in a few locations due to special infrastructure requirements, such as space, power, and cooling, and the need for a qualified workforce and associated management costs [18]. However, cloud users are scattered worldwide, and consequently, many users are not geographically close to the cloud data centers of their preferred cloud provider. In the same way, user devices are scattered. They may also be distant from the cloud, resulting in unacceptable access delays for various applications (e.g., interactive and virtual/augmented reality). The computing continuum will transparently offload data, process application components, and store data in distant facilities whenever application requirements allow, but will also use surrounding facilities, devices, and networking to provide low latency access to data and processing capabilities.

2.1.6. Serverless computing

Serverless computing emerged in the mid 2010's, becoming popular in 2016. It refers to hiding infrastructural details in cloud computing, primarily in terms of instantiation and autoscaling, along with fine-grained billing of mostly stateless compute units, so-called functions, or containers with function semantics. Thus, serverless computing is largely seen as an evolution of cloud computing towards a broader circle of software-as-a-service engineers. A serverless application is typically composed of several functions coupled with low-latency backend services. The challenges related to infrastructure hiding relate to more unpredictable service response times due to effects such as cold start and spawn start, as well as difficulties in constrained deployments that require service discovery and failure recovery [19].

As in most cloud computing approaches, serverless assumes a centralized execution model. However, the first approaches exist to adjust this paradigm to previous concepts around distributed cloud computing, effectively leading to distributed serverless capabilities. The computing continuum is expected to evolve serverless computing by allowing developers to focus on application logic and requirements, while decision-making on execution and data placement to support heterogeneous serverless requirements will inherently support the continuum infrastructure.

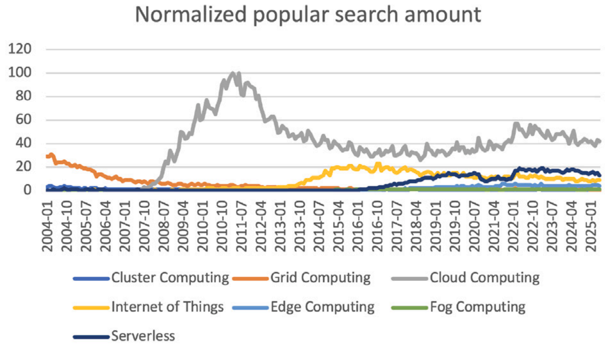

Fig. 2. Search popularity in the Google search engine for different distributed computing paradigms over the years.

2.1.7. *Fog, edge and mobile computing*

Fog computing and edge computing are distributed computing paradigms that started to appear around 2012 [6] to support IoT applications and evolved during the past decade [20]. To illustrate the timeline of these technologies, Fig. 2 illustrates the popular search for each term from 2004 until today, with data generated by Google Trends tool.[1]

While the end devices are concentrated at the network edge, the fog devices are distributed from the edge access points through the network core [3]. The cloud is farther away, requiring requests from the edge to traverse the public Internet to access cloud computing resources. As opposed to Content Delivery Networks (CDNs) [21], which are focused on distributing data storage generated at centralized servers, edge and fog are also concerned with data processing closer to the end user, where data is often generated. Cloud, fog, and edge computing can work in synergy to support heterogeneous IoT applications, thus constituting a multi-layered infrastructure. As the fog infrastructure can be composed of different levels [7] as the mid-layer of the Edge–Fog–Cloud infrastructure, this composition can offer a variety of quality of service levels [22]. Fog computing and Multi-access Edge Computing (MEC) are closely related concepts: Fog Computing focuses on the Internet of Things (IoT) while MEC focuses on the mobility of devices in the network. MEC aims to deploy services close to mobile end-users to reduce latency, while Fog aims to provide computing infrastructure between the edge and the cloud. MEC follows guidelines established by the European Telecommunications Standards Institute (ETSI) Network Function Virtualization (NFV) Management and Orchestration (MANO) [23], being more closely related to network and services management. In contrast, Fog computing follows architectural principles established by the ETSI Machine-to-Machine (M2M) technical committee, more closely related to computing infrastructure and offloading to support data generated by devices.

Due to the hierarchical architecture [24], bi-directional communications between the fog and cloud are essential in fog computing. For instance, a service presenting high computational requirements is deployed in the cloud, needing to communicate with another service placed in the fog to reduce data transportation to the cloud. These interactions must be considered in the allocation process, leading to complex service dependencies that must be guaranteed. Depending on their different requirements, applications can be deployed and run on any device in this infrastructure composition. Moreover, depending on the application's needs and the user's geographical location, components can be distributed among devices at different levels [25]. Devices that connect to the Fog–Cloud computing infrastructure often do so through wireless connections (e.g., 5G/6G and WiFi), thus also depending on the availability of communication channels that can fulfill their requirements.

The distribution (e.g., density or number of levels) of the Fog–Cloud hierarchy can vary from place to place. Still, the first level is expected to be located one hop away from the edge (user or device): at the access point (e.g., WiFi or cell phone antennas) or immediately above it [3]. This would be the first (closest) offload option for devices at the edge, providing lower latencies. However, likely with limited computing capacity, but when combined with the cloud, it can provide the necessary computing power for applications with heterogeneous requirements. Additional fog levels can be added to enhance computing capacity closer to the edge, according to infrastructure providers' demand and capacity planning.

It is common to designate the aforementioned hierarchy of computing capacity as *fog nodes*, *cloudlets*, or *micro data centers* [26]. Conceptually, the higher in the hierarchy a cloudlet is, the larger its processing/storage capacity since it is expected to support more devices in the tree downwards the edge. In contrast, cloudlets that are higher in the hierarchy are also expected to present longer network delays to the client or data consumer at the edge. Therefore, the hierarchical composition of micro data centers (or cloudlets) with the cloud provides a range of computing capacity at different geographical (and logical) distances to the devices at the edge.

The IoT–edge/fog–cloud infrastructure currently leads to a preview of a canonical computing continuum. Still, it does not allow interoperable, transparent, and autonomous management of resources, data, and applications across computing and networking devices.

2.1.8. *In-transit computing*

In-transit computing can be seen as an intermediate solution towards the continuum, where delays are masked by processing data in equipment placed throughout the network between the edge and the core [27]. This type of processing is often associated with stream processing and complex event processing, where in-transit computing equipment can be used to rapidly run filters, aggregation functions, and other small/fast processing functions. However, limits in processing capacity within the network provide low reliability in terms of performance guarantees. In addition, a transparent continuum of computing is expected in the future, where propagation delays dominate data transfer times and the data transfer and application processing are jointly managed, composing a unique computing and networking infrastructure to support the requirements of any set of applications.

2.2. *Present: The computing continuum literature*

Current literature often defines the computing continuum as a composition of existing infrastructures, as further discussed in the review presented in this section. Fig. 3 illustrates this view, which brings many new challenges to resource allocation and management [3], but falls short in actually providing a computing continuum as there is a clear separation of terminologies, management mechanisms, computation and communication infrastructures, devices classes, and enabling technologies at the different layers. Moreover, current networking limitations allow us to differentiate the hierarchy levels based on bandwidth and access times, creating different infrastructure classes.

2.2.1. *Literature overview*

This section reviews the current literature that touches on the computing continuum in different ways, classifying papers concerning the computing continuum terminology, the considered computing/networking infrastructure and motivation, the main focus of each paper, and how the evaluation is presented in each paper. Before entering the related work analysis, we present in Table 1 the characteristics used to classify the papers.

[1] https://trends.google.com/. According to a disclaimer in the Google Trend platform, the spike around 2022-01 indicates an improvement in the data collection system for the google trend tool.

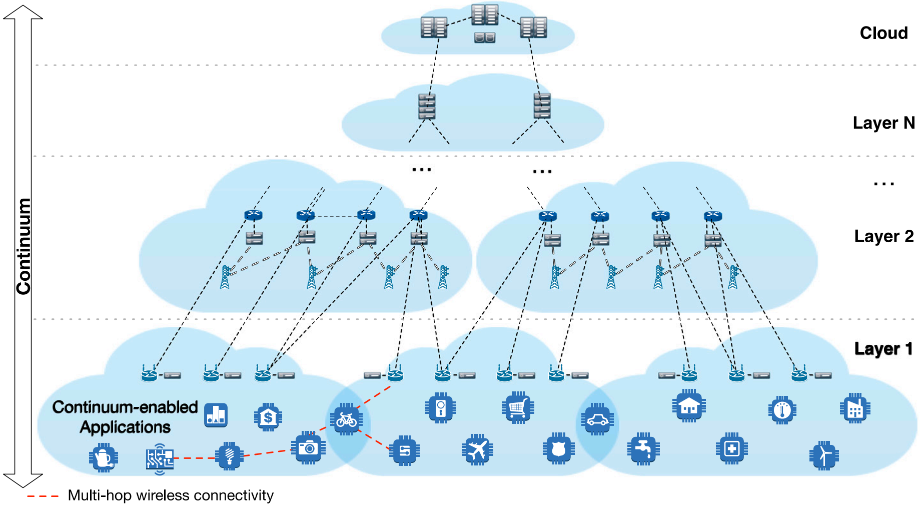
Cloud
Layer N
...
Layer 2
Layer 1
Continuum
Continuum-enabled Applications
Multi-hop wireless connectivity

Fig. 3. Today, the Computing Continuum can be seen as a composition of resources from the Edge to the Cloud. *Source:* Adapted from [3].

Table 1
Characteristics used to classify the literature.

Main terminology	This characteristic separates the paper into different terminologies considered in the reviewed paper, namely **Cloud Continuum**, **Computing Continuum**, and **X-to-Cloud Continuum**, where *X* commonly includes words such *Fog*, *Edge*, and *IoT*, and sometimes more than one word, as for example in *IoT–Edge–Cloud Continuum*.
Infrastructure/ Motivation	This characteristic separates the reviewed paper in terms of the considered infrastructure. A paper is in the **Network and 5G/6G** category when the main motivation or analysis is concerned with networking aspects, including cellular communications. On the other hand, the remaining categories, namely **Cloud**, **Fog**, **Edge**, **Mobile Edge**, and **IoT**, are more concerned with the computing infrastructure aspects of the continuum in different distributed architectures and technologies.
Focus	This characteristic classifies the papers into different focuses. **Architecture or Framework** papers proposed and describe conceptual designs of the continuum. **Management and orchestration** concentrates in the general aspects of infrastructure management and data processing orchestration. **Resource allocation and optimization** papers are focused on discussing and proposing resource allocation techniques and optimizing the use of the continuum infrastructure. **AI/ML** papers propose and analyze the use of artificial intelligence and/or machine learning techniques in any applicable aspects in a computing continuum scenario. Similarly, **Security and Privacy** papers are concerned with these aspects in a computing continuum scenario. Papers in the **Tool/Prototype** category focus on specific tools that can be applied in the computing continuum, and/or use real, small scale deployments (prototypes) to illustrate the continuum concept. In addition, papers can have focus on specific **Applications**.
Evaluation	This characteristic lists how the reviewed literature discusses/evaluates the continuum context presented in each paper. A **Review/Tutorial/Definition** paper provides a more conceptual discussion over the literature and the continuum aspects. A **Theoretical** evaluation provides formal theoretical analyses/results for a continuum context. **Simulation** papers use simulation tools to analyze performance aspects, thus not relying on real devices but on simulated environments, while **Testbed** papers use testbeds in the evaluation analyses. These testbeds are usually either based on real heterogeneous devices (cloud/edge servers, Raspberries/Arduinos, etc.) or on configurable virtualized environments. Finally, **Use case/ Proof of Concept** papers present deployed evaluations that try to mimic characteristics of the continuum in small scale.

Table 2 details these characteristics of the related work. We searched for the pattern *("Computing Continuum" OR "Cloud continuum")* in Google Scholar[2] first in February 2024 and then in April 2025. In the first search, we collected the list of the first 100 papers published between January 2020 and February 2024 as sorted by the Google platform. The inclusion criterion was based solely on the publisher, where we included papers published by IEEE, ACM, Springer Nature, Elsevier, and Wiley. We excluded other publishers, non-peer reviewed, and pre-print articles, thesis, and reports. We chose not to consider citation amount as a criterion to eliminate the most recent papers relevant to this study. After the first search and selection based on this inclusion criterion, 78 papers were analyzed. Then, in the second search, we used the same protocol, but now with papers published solely in 2024 and 2025, and selected the first 22 papers that did not overlap with the first search, resulting in 100 papers. We also included four papers suggested to us by reviewers of the paper. It is important to note that each column of the table touches an aspect of the paper, as defined in Table 1, but each paper approaches each of these aspects with different depths. Thus, naturally, a paper that is classified in several columns may have a different balance of depth in each covered topic when compared with other papers with similar classifications.

From Table 2, we observe that most papers reviewed adopt as the main Continuum terminology the term"X-to-Cloud Continuum" (64 papers), while "Computing Continuum" is adopted by 38 papers. When considering the main infrastructure/motivating aspect, most works naturally focus on the Cloud and Edge, as the Continuum is currently seen as an extension of the Cloud towards the Edge of the network. We observe that even if the main used terminology is the *Computing Continuum* without including any Cloud wording (as in Edge-to-Cloud Continuum), 75 research papers consider cloud computing as the main infrastructure supporting the computing continuum. On the other hand, including the network technologies and mobile devices in the continuum has been less explored in the literature (20 and 19 papers, respectively), highlighting the significant potential for future research and development in these areas. Therefore, the interplay between networking, mobile devices, and computing infrastructure is a central challenge that needs to be explored to develop the computing continuum to its full extent.

[2] http://scholar.google.com.

Regarding the focus of the reviewed literature papers, 48 manuscripts are concerned with presenting and discussing architectures or frameworks, which is natural when an emerging distributed infrastructure is being discussed in the literature. Resource allocation and optimization, and resource management and orchestration are also studied in a relevant number of papers: 53 and 50, respectively. However, security and privacy issues and mechanisms still need to be thoroughly addressed in the reviewed literature, with only 11 papers focusing on these aspects.

Concerning the kind of evaluation presented in each paper, we observe that testbeds are the most common, with 45 papers using continuum-like testbeds, even though they are set up as a composition of IoT–edge/fog–cloud infrastructures on a small scale. A computing continuum testbed comprising several devices on a large scale has yet to be deployed. Another 25 papers focus on reviews, tutorials, or definitions; 28 use simulation as the evaluation tool; and 27 present use cases or proofs of concept of the computing continuum or its components.

We highlight some correlations that we identified by analyzing Table 2 in more detail, suggesting that these characteristic appear together: 6 of 20 papers that are motivated by network and 5G/6G are also concerned with MEC infrastructures; 20 of 48 papers with a focus on architecture and frameworks are evaluated through use cases/proof of concepts; 32 of 53 papers that focus on management and orchestration are evaluated using testbeds; 24 papers focused on resource allocation and optimization are evaluated using simulations; and 13 of 19 papers that MEC in the computing infrastructure are also focused on resource allocation and optimization. Management and orchestration is also often the focus of papers that consider IoT (23 out of 44) and edge (40 of 66). Moreover, all papers that have theoretical results also use simulations in their evaluation.

The literature review shows us that most authors consider the computing continuum a straightforward composition of current paradigms such as IoT, edge/fog, and cloud. In this paper, we aim to evolve this understanding, moving this composition forward to a seamless distributed system where computing capacity is, in fact, part of a continuum that encompasses data collection, processing devices, and networking devices and where resource management is transparent and jointly performed by the network and the computing infrastructures altogether and cooperatively.

2.2.2. Related work

In the revised literature, we identified 25 papers focusing on reviews, tutorials, or definitions of the computing continuum. Differently from this paper, various review papers focus on specific aspects that should be considered to advance towards the continuum: testbeds [128], energy consumption [117], monitoring [101], learning and reasoning [43,56] and distributed/edge intelligence [73,94,118], big data [72], robotic systems support [79], programmability [29], resource management [46,106,113], software architecture [103], urgent science [28], in-network computing [51], military networks [71], and real-time learning for vehicular computing [45].

These reviews illustrate the broad applicability of the continuum and the specifics that must be considered in a single computing framework that allows the management of heterogeneous computing devices and networking.

The seven other literature review papers found discuss broader aspects of the computing continuum. Pujol et al. [55] briefly discuss four research lines that will be important for the computing continuum, namely self-adaptation, inter-relations, and monitoring & knowledge, connecting the idea of the computing continuum to complex systems such as the human body. Dustdar, Pujol, and Donta [57] evoke the Cloud–Fog–Edge–IoT infrastructure as composing the multiple tiers of the computing continuum. The paper argues that current methodologies will be inappropriate for the continuum. It discusses a new methodology based on the Markov Blanket, stating it provides a more flexible management framework and brings some similarities to Cloud Computing systems management. However, the paper does not include the network as an integral part of the computing continuum that should be integrated into the computing management framework to provide a seamless continuum. Moreover, the paper does not review the literature on the computing continuum, even though it discusses existing general approaches and learning mechanisms that could help manage the computing continuum. Milojicic [30] reports on a virtual round-table with three panelists, namely Tom Bradicich, Adam Drobot, and Ada Gavrilovska, who present their views on aspects such as the future of computing, market forces driving the continuum of computing, key market verticals/industries that suitable for the continuum, key programming paradigms for edge-to-cloud technologies, end-to-end feasibility for the cybersecurity-trust-privacy triplet, and the role of virtualization and AI in edge–cloud setups towards the continuum. It also touches on the importance of 5G for edge–cloud deployment, suggesting the composition of networking and computing infrastructures, as we also advocate in this paper. Finally, the panelists put their views on the foreseen applications in the computing continuum. Khalyeyev, Bureš, and Hnětynka [60] review the literature on the edge–cloud continuum, emphasizing expected features and discussing challenges. The authors organize existing views of the edge–cloud continuum, showing heterogeneity of computing resources and arguing that network infrastructures have a vital role in interconnecting these computing resources; also, a computing hierarchy can result from categorizing entities in the edge–cloud continuum. The paper concludes that a hierarchical structure, componentized development, and dynamic behavior of software components, liquid and adaptable software, and edge intelligence are the key properties of edge–cloud infrastructures. Adyson Maia et al. [113] offer a comprehensive survey of current research focused on the seamless and efficient delivery of communication, computing, and caching (3C) services within the cloud-to-edge continuum. It highlights AI-driven and collaborative solutions aimed at optimizing this process. The study examines the interplay between communication, computation, and caching by analyzing key use cases, emphasizing their combined importance in supporting next-generation network infrastructures (NGNIs). The authors explore the opportunities and challenges of integrating 3C resources into NGNIs. They provide insights into architectural considerations, regulatory implications, and business perspectives. Auday Al-Dulaimy et al. [104] discuss computing models with a focus on cloud computing, the computing models that emerged beyond the cloud, and the communication technologies that enable computing in the continuum. The authors propose two reference architectures: one for edge–cloud computing models and another for edge–cloud communication technologies. To validate these architectures, they examine real-world use cases from various domains, including industry and scientific research, and demonstrate how they align with the proposed models. Finally, Serpanos [130] broadly discusses trends and challenges of the computing continuum, focusing on application scenarios and the role of AI, and discussing resource allocation challenges in the composition of resources from the edge to the cloud.

This paper brings a broader literature review and provides a conceptual view of the continuum that differs from the review papers found in the literature, emphasizing the need for resource and application management frameworks that encompass computing and networking technologies to support adaptable applications that are technology agnostic and can find and process data anywhere in the continuum without specific management or development efforts. In the next section, we describe our vision and definition of the computing continuum, discussing the distributed computing infrastructures and networking evolution moving towards a computing continuum.

Table 2
Characteristics from the literature related to ''Cloud Continuum'' and ''Computing Continuum'' between January 2020 and April 2025 [28–131].

Year	Ref.	Main continuum terminology			Infrastructure/motivation						Focus							Evaluation					Ref.	Year
		Cloud Continuum	X-to-cloud continuum	Computing Continuum	Network and 5G–6G	Cloud	Fog	Edge	Mobile Edge	IoT	Architecture or Framework	Management and orchestration	Resource allocation and optimization	AI/ML	Security and Privacy	Tool/ Prototype	Application	Review/ Tutorial/ Definition	Theoretical	Simulation	Testbed	Use case/ Proof of concept		
2020	[28]			•	•			•			•						•	•					[28]	2020
2020	[29]			•		•	•	•		•	•					•	•	•					[29]	2020
2020	[30]		•						•			•		•		•	•	•					[30]	2020
2020	[31]			•		•		•		•	•	•	•			•					•	•	[31]	2020
2021	[32]		•			•		•		•	•	•								•			[32]	2021
2021	[33]		•	•	•	•	•	•	•							•	•					•	[33]	2021
2021	[34]			•		•	•								•								[34]	2021
2021	[35]		•	•		•		•			•		•			•	•				•	•	[35]	2021
2021	[36]		•			•			•	•			•	•			•			•			[36]	2021
2021	[37]		•			•			•		•		•				•					•	[37]	2021
2021	[38]		•			•		•					•	•							•		[38]	2021
2021	[39]			•							•		•								•		[39]	2021
2021	[40]		•			•		•		•	•	•	•								•		[40]	2021
2021	[41]		•		•	•				•	•	•				•					•	•	[41]	2021
2021	[42]		•			•						•		•								•	[42]	2021
2021	[43]		•	•										•				•					[43]	2021
2021	[44]		•			•	•	•					•	•						•			[44]	2021
2021	[45]				•			•	•	•	•	•		•				•					[45]	2021
2021	[46]			•		•	•	•	•	•		•		•				•				•	[46]	2021
2021	[47]	•				•		•			•	•				•						•	[47]	2021
2021	[48]			•							•											•	[48]	2021
2021	[49]		•	•		•				•		•	•			•					•		[49]	2021
2021	[50]		•			•	•	•			•	•				•					•	•	[50]	2021
2021	[51]		•		•					•		•	•	•			•	•					[51]	2021
2021	[52]		•			•		•						•			•				•		[52]	2021
2022	[53]			•		•	•				•		•				•				•		[53]	2022
2022	[54]		•								•	•					•				•	•	[54]	2022
2022	[55]			•									•	•	•			•					[55]	2022
2022	[56]			•		•	•	•		•	•	•		•				•					[56]	2022
2022	[57]			•		•	•	•		•	•	•		•				•				•	[57]	2022
2022	[58]		•			•			•				•							•			[58]	2022
2022	[59]		•		•	•		•	•	•			•						•	•			[59]	2022
2022	[60]		•						•	•								•					[60]	2022
2022	[61]			•						•	•	•		•			•				•	•	[61]	2022
2022	[62]		•										•	•						•	•		[62]	2022
2022	[63]		•			•			•		•		•	•					•	•	•		[63]	2022
2022	[64]		•			•			•		•		•	•						•			[64]	2022
2022	[65]			•		•	•	•					•							•	•		[65]	2022
2022	[66]		•			•	•			•			•				•			•			[66]	2022
2022	[67]		•		•								•		•					•			[67]	2022
2022	[68]		•			•	•	•				•									•	•	[68]	2022
2022	[69]		•			•			•			•	•								•	•	[69]	2022
2022	[70]		•			•		•			•	•									•		[70]	2022
2022	[71]	•			•		•	•									•	•					[71]	2022
2022	[72]			•							•							•				•	[72]	2022
2022	[73]		•	•		•		•	•					•				•					[73]	2022
2022	[74]		•			•		•		•	•	•	•			•				•		•	[74]	2022
2022	[75]		•		•	•		•				•					•				•		[75]	2022
2022	[76]			•							•	•	•	•		•					•	•	[76]	2022
2022	[77]		•			•		•				•	•			•	•				•		[77]	2022
2022	[78]			•	•							•				•	•			•	•		[78]	2022
2022	[79]	•	•			•		•								•	•	•					[79]	2022
2023	[80]		•		•	•		•			•	•								•	•		[80]	2023
2023	[81]			•			•						•				•			•			[81]	2023
2023	[82]			•		•		•			•	•				•					•	•	[82]	2023
2023	[83]		•			•			•				•						•	•			[83]	2023
2023	[84]		•		•	•			•				•						•	•			[84]	2023
2023	[85]		•			•		•		•	•	•				•	•				•		[85]	2023
2023	[86]			•				•		•		•									•		[86]	2023
2023	[87]		•			•		•			•					•						•	[87]	2023
2023	[88]		•			•		•					•	•					•	•			[88]	2023
2023	[89]		•		•	•		•					•						•	•			[89]	2023
2023	[90]		•			•		•			•	•									•		[90]	2023
2023	[91]		•								•										•		[91]	2023
2023	[92]		•			•		•			•				•						•		[92]	2023
2023	[93]			•								•	•				•					•	[93]	2023
2023	[94]			•		•		•		•		•		•				•					[94]	2023
2023	[95]		•			•		•				•	•							•	•		[95]	2023
2023	[96]		•			•		•	•			•	•							•	•		[96]	2023
2023	[97]		•			•		•		•		•	•								•		[97]	2023
2023	[98]		•			•		•			•	•	•								•		[98]	2023
2023	[99]		•		•								•							•			[99]	2023
2023	[100]			•						•	•					•						•	[100]	2023
2023	[101]	•				•										•		•					[101]	2023
2023	[102]			•		•	•	•			•	•									•		[102]	2023
2024	[103]		•		•			•				•						•			•		[103]	2024
2024	[104]			•	•	•		•		•	•	•	•				•	•		•			[104]	2024
2024	[105]			•		•	•	•				•				•					•		[105]	2024
2024	[106]			•		•	•	•		•		•		•				•			•		[106]	2024
2024	[107]		•		•	•		•			•	•			•								[107]	2024
2024	[108]			•		•		•		•	•	•	•	•							•		[108]	2024
2024	[109]		•			•		•		•	•	•			•								[109]	2024
2024	[110]		•			•		•			•	•											[110]	2024
2024	[111]	•				•		•		•							•				•		[111]	2024
2024	[112]		•			•	•			•			•										[112]	2024
2024	[113]		•		•	•		•		•		•	•	•	•			•					[113]	2024
2024	[114]		•			•		•			•	•				•					•	•	[114]	2024
2024	[115]		•		•				•				•	•									[115]	2024
2024	[116]			•		•		•			•		•						•	•			[116]	2024
2024	[117]	•				•	•	•			•	•						•				•	[117]	2024
2024	[118]			•				•						•	•			•					[118]	2024
2024	[119]		•			•		•			•		•									•	[119]	2024
2024	[120]		•					•						•			•				•		[120]	2024
2024	[121]		•			•		•	•			•	•								•		[121]	2024
2024	[122]		•		•				•		•	•	•				•				•		[122]	2024
2024	[123]		•			•		•			•		•							•			[123]	2024
2024	[124]		•							•	•			•			•			•			[124]	2024
2024	[125]			•		•		•			•		•	•								•	[125]	2024
2024	[126]			•		•	•	•		•			•		•					•			[126]	2024
2025	[127]		•			•		•				•	•	•			•					•	[127]	2025
2025	[128]	•				•	•	•		•		•	•		•			•		•			[128]	2025
2025	[129]			•		•		•		•	•	•			•						•		[129]	2025
2025	[130]			•		•		•						•			•	•					[130]	2025
2025	[131]		•			•		•		•		•	•			•					•		[131]	2025

3. Defining the computing continuum

The computing continuum can be seen as a generalized model for the hierarchy presented in Fig. 3, where edge, fog, and cloud computing are merged into a seamless continuum of computing and networking capacity that applications can transparently utilize.

Sidney Karin and Susan Graham articulated an early vision for a continuous distributed system in 1998 [132]:

> *"... the emergence of the World-Wide Web has produced another trend – the dynamic use of multiple physical computer systems viewed by the end user as a single system. Such systems might be considered continuous."*

The above vision relates to the pervasive and ubiquitous computing we have experienced with the popularization of the Internet [133]. In addition to cloud continuum definitions [134], which claim an extension of the cloud to scattered devices, in our context we are also interested in a more generalized vision as an extension of the pervasive and ubiquitous computing paradigm. Data communication speed is increasing in the computing continuum devised in this paper. It can be virtually unbounded, turning the "pervasive continuum" system into a **space–time** continuum: data location in the distributed system (**space**) and processing speed (**time**) shall be combined into a single system component (the continuum) as delays for both data transfer and processing are often constrained to the same magnitude regardless of the data size. Based on the discussion above, our definition of a computing continuum is as follows.

> *The computing continuum is a highly heterogeneous and adaptable distributed system capable of providing seamless data transfer and computation offloading under feasible quality of service requirements. The computing continuum spans computational resources scattered from the edge to the core of a network, composing an infrastructure capable of supporting delay requirements asymptotically bounded by physical distances.*

In contrast, the literature presents technology-attached definitions, as below:

> *"Cloud Continuum is an extension of the traditional Cloud towards multiple entities (e.g., Edge, Fog, IoT) that provide analysis, processing, storage, and data generation capabilities"* [134]

> *"Ultimately, new applications are emerging that take advantage of all the computing tiers available, hence, they are systems that are simultaneously executed on the Edge, Fog, and Cloud computing tiers. These systems are known as computing continuum systems".* [135]

Today, as discussed in the technology-attached continuum definitions [3,134,135] and illustrated in Fig. 3, an instance of the computing continuum brings edge resources that include mobile phones and IoT devices, while core network resources include cloud datacenters. As the next step in the evolution toward the computing continuum we devise, Content Delivery Networks will be complemented and enhanced by the so-called Cloudlets or microdatacenters, in-between the edge and the core, and by in-network computing devices to be largely deployed across 5G and 6G network equipment. Eventually, the computing continuum of all these computing and networking resources will allow a technology-agnostic deployment for seamless data transfer and computing offloading. Distributed computing models in the continuum have yet to be proposed, and with this paper, we aim to provide a basis for the development of such a paradigm, leading to new resource management models for distributed systems.

Fig. 4 illustrates the past and future of *access to* distributed computing infrastructures to characterize such an evolution in distributed computing systems. It shows a timeline from past systems (e.g., mainframes) to the future Computing Continuum. In addition, from a networking perspective, remotely accessing a mainframe in the past would result in high response times, as additional hops between the user and the mainframe added significant delay because of previous limitations in network equipment processing power and bandwidth.

When cluster computing emerged, remote access to distant cluster nodes would also be impacted by network equipment delays at each hop. Still, nowadays access to Cloud Computing facilities is prone to bandwidth limitations, but networking equipment and bandwidth evolution allow faster networking processing, resulting in significantly smaller delay enlargement when more hops are introduced. In the past, an important part of the delay to remote access clusters came from propagation latencies in the communication media and bandwidth bottlenecks.

The *canonic* computing continuum comes from an indefinite increase in network performance, in which network equipment introduces infinitesimal, negligible delays, resulting in a scenario where network delays become continuous as propagation latencies in the communication medium dominate the total delay. This trend means that more hops of communication do not result in leaps in delays introduced by network equipment queueing, processing, and transmission, as would be expected in previous distributed systems. The computing continuum is characterized by increased data transmission rates such that delays introduced by the network become less and less apparent. Therefore, in a full computing continuum, the physical medium speed barriers impose most of the constraints for data transfers and request–response times. This is conceptually illustrated in Fig. 4 by the concept of quantum communications, where the computing continuum expected to be deployed over previous network technologies could be explored to its full extent.

According to Foster [136], the performance landscape changes with the computing continuum, and response time bounds could depend more on a distance radius than on the network topology as the computing power is scattered and the network speed becomes unbounded. Therefore, the evolution to the continuum is characterized by a reduction in the importance of delays other than propagation. Kurose and Ross [137] identify four sources of packet delays that make up a nodal delay, i.e., the delay introduced by a node (e.g., a router or switch) in the network path:

$$d_{nodal} = d_{proc} + d_{queue} + d_{trans} + d_{prop}, \tag{1}$$

where

- d_{proc} is the processing time of a packet in the given node
- d_{queue} is the time a packet spends in the queue of the node
- d_{trans} is the time taken to push a packet into the communication channel
- d_{prop} is the time the bits take to propagate in the communication media.

The evolution towards the computing continuum will decrease d_{proc} and d_{queue} as the processing capacity increases in the network equipment, while d_{trans} will be reduced due to the increase in bandwidth. In contrast, d_{prop} should not change, as propagation delays are imposed by physics in the channel. Therefore, delays in the continuum would approximate propagation delays as the bandwidth increases (Eq. (2)).

$$d_{continuum} \simeq d_{prop} \tag{2}$$

The delay in the continuum is lower bounded yet is approximated by the propagation delay, regardless of the number of hops between the client and the computing facilities fulfilling the application requirements. As a result, we argue that distributed systems must incorporate geographical distances, reflected as propagation delays, as parameters in resource management to optimize the Quality of Service (QoS) of applications and keep users' Quality of Experience (QoE) at acceptable

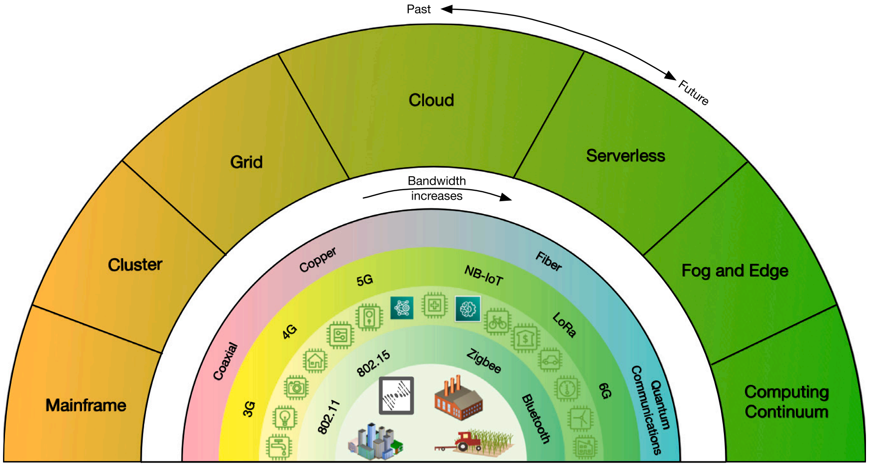

Fig. 4. The illustration of distributed computing infrastructures from a **historical perspective**. Access times in older systems were higher due to leaps in delays introduced by networking equipment and low bandwidth communication channels. In the future, these leaps will be less noticeable as networking equipment and communication medium performances increase, forming the Computing Continuum as a composition of resources that seamlessly access with low-to-no network penalties.

Table 3
Distributed computing infrastructures in an **updated, snapshot perspective**: how each type of system currently presents itself to the users in terms of access latency, location of the infrastructure, computational and storage capacity, what is the cost to access each infrastructure, connectivity/reliability, and if the expected performance guarantees for each type of system.

		Cluster	Grid	Cloud	Serverless	Fog	Edge	Ubiquitous	In-transit	Continuum
User access latency		Low	High	High	High	Low	Low	Low	Medium	Low
Location	Core			✓	✓	✓				✓
	Edge	✓		✓	✓		✓	✓		
	Distributed edge		✓			✓	✓			✓
	Along data path								✓	✓
Computation/Storage capacity		Medium	High	High	High	Medium	Low	Low	Low	High
Data access cost		Low	High	High	High	Low	Low	Medium	Medium	Low
Connectivity/reliability		High	Medium	High	High	Medium	Low	Low	High	High
Performance guarantees		High	Low	Medium	Medium	Medium	Low	Low	Low	High

levels. Moreover, besides considering propagation delays as a core application optimization criterion, incorporating energy consumption in the decision-making equation can help reduce the carbon footprint of computing systems supported by the continuum [138].

Table 3 illustrates a set of characteristics of distributed processing systems nowadays. This table focuses on a snapshot of how each type of system currently presents itself to the users, thus differing from the illustrative evolutionary perspective shown in Fig. 4. An essential difference from the historical viewpoint is that access **latencies** to many computing clusters are nowadays lower due to higher capacity links connecting users to the cluster computing infrastructures often located at the edge, nearby (e.g., within the same city) facilities. On the other hand, Grid and Cloud computing usually present higher **access latencies**, as grids are geographically scattered and (public) cloud data centers are frequently **located** thousands of kilometers away from the users. Private clouds can be **located** far from their users when leased from cloud providers or large multinational corporations. This results in **lower data access costs** for clusters compared to grids and clouds. Moreover, compared to grids, clusters, and clouds usually have **more reliable communication networks** among their resources and also in their connections to ISPs or backbones, while grids can encompass heterogeneous and **less reliable resources** at different tiers, reducing grid's **performance and data transfer times guarantees**. On the other hand, the controlled resource management in cluster environments can improve **performance guarantees**. In contrast, cloud resource sharing and public communication channel access can impair producing reliable **performance guarantee** levels.

While serverless computing derives from the cloud business model, inheriting many of the cloud computing characteristics, fog, and ubiquitous computing consider computing resources are **widely distributed**. Ubiquitous (pervasive) computing can be seen as an edge, device-based computing resembling current Internet of Things definitions with no additional computing infrastructure. On the other hand, fog computing is more concerned with deploying a computing environment complementary to the cloud but closer to those pervasive computing devices, thus creating a **distributed edge infrastructure** toward the cloud data centers. Consequently, it **lowers data access costs** in fog as it aims to keep data closer to where it is needed, while ubiquitous computing relies on accessing data from devices that can be anywhere. Lastly, in-transit computing aims to process data during communication, i.e., in transit. This computing in the network concept aims to perform partial processing of data to improve equipment utilization and speed up computing results and application response times. Attaching fog computing nodes (e.g., cloudlets/micro data centers) to access points and at some critical points in the core network would also be complementary to this approach.

The computing continuum embraces all of the above by creating seamless connectivity and incorporating the characteristics of these computing infrastructures. As computer networks evolve towards faster and more reliable connectivity, the computing continuum will present the best attributes of these systems altogether. Notwithstanding, distributed systems and networking challenges of all kinds emerge to enable this seamless connectivity to be exposed to application developers, system managers, and final users.

4. Application scenarios and impact

The emerging continuum infrastructure and novel technologies are leading to new use cases requiring even more stringent requirements,

Table 4
Expected requirements for emerging use cases [11,139–141].

Use case	Latency	Reliability	Throughput
Self-driving vehicles	<1 ms	99.9999999%	1–10 Mbps
Extended reality (XR)	<1 ms	99.99999%	>1 Tbps
Holographic streaming	<1 ms	99.99999%	>1 Tbps
Tactile internet	1–10	99.999%	1–100 Mbps
Industrial IoT	1–10 ms	99.9999999%	1–10 Mbps

such as ultra-bandwidth data rates and ultra-low latency communications. This section presents insights into some of the envisioned application scenarios that will arrive with the 6G continuum [139]. Although the computing continuum is paradigm-agnostic for networking technologies, the 6G continuum is the first personification of the computing continuum. It presents an opportunity to integrate network and processing management toward realizing the computing continuum. In this section, we embrace this scenario to illustrate application scenarios and the impact of the computing continuum.

During its deployment, 5G presented trade-offs between different factors, including latency, energy efficiency, deployment costs, throughput, and reliability, meaning that different deployment configurations of 5G networks are typically instantiated to accomplish various use cases (e.g., high-bandwidth versus low-latency communications). In contrast, the 6G continuum will possess computing resources throughout the network area to jointly support the highly demanding requirements of the foreseen applications, as shown in Table 4. The following sections discuss the main characteristics and the foreseen research challenges of these emerging use cases in light of the envisioned 6G continuum.

4.1. Highly-mobile self-driving vehicles

Self-driving or autonomous driving means that cars or other vehicles, such as electric vertical take-off and landing (eVTOL) aircraft [142], can navigate with little or no human intervention. These vehicles rely on sensing the environment to enable software-based driving decisions. Fully autonomous transportation systems will offer safer traveling, improved traffic management, and support for infotainment in an intelligent transportation system, also known as the Internet of Vehicles (IoV) [143]. These use cases where moving vehicles are part of the computing continuum demand unprecedented levels of reliability, higher than nine nines (i.e., 99.9999999%), with ultra-low latency communications below 1 ms. This becomes even more challenging for high-mobility scenarios (i.e., up to 1000 km/h), such as futuristic transportation use cases, including Hyperloop or other tube-based transport systems, and high-speed aerial vehicles or eVTOLs connected to non-stationary satellites. These scenarios pose unique challenges and opportunities for networked systems, particularly in the context of 5G/6G connectivity in the computing continuum. Combining such stringent requirements with intelligent and collaborative vehicular decision-making guarantees passengers' and pedestrians' safety [144], which is challenging to achieve in today's networks.

In high-mobility environments, the user or device frequently transitions between network cells, requiring seamless handovers. The latency and reliability implications are critical, especially for safety-critical applications. Solutions must include fast and predictive handover mechanisms, perhaps aided by machine learning and low-latency edge computing. Also, because of variable signal strength and rapid location changes, data offloading between edge and cloud must adapt dynamically. For example, edge resources near highways or railways can cache content or pre-process data to reduce round-trip delays. Mobility-aware orchestration mechanisms are essential to ensure continuity of service. High-mobility use cases will also benefit from precise and low-latency localization. In turn, accurate positioning can improve network performance by enabling beamforming, route prediction, and context-based resource allocation. This is critical for applications like fleet coordination, urban air mobility, or VR streaming in transit.

In addition, the high number of sensors per vehicle will generate an enormous amount of data that needs to be processed at the closest proximity of vehicles since these vehicles will need to make decisions quickly to respond to the driving challenges they face. In addition, Unmanned Aerial Vehicles (UAVs) will open possibilities for the development of further applications (e.g., emergency response services and video surveillance) due to the increased network connectivity [145]. The 6G continuum will pave the way for connected vehicles through advances in hardware and software and the possibility of running software at all levels of the infrastructure.

4.2. Bandwidth-intensive Extend Reality (XR) applications and the metaverse

Extend Reality (XR) represents one of the most data-hungry applications of 6G networks, requiring up to 2.3 Tbps for high-quality streaming with end-to-end latencies lower than 10 ms [146]. It can be defined as all real and virtual environments created via computers or wearables that blend virtual and physical worlds to create fully immersive experiences. XR will completely revolutionize multimedia service delivery. These applications require sub-millisecond latency while their throughput requirements exceed 1 Tbps since data cannot be compressed to attain fully-immersive experiences in real-time.

Furthermore, the envisioned XR applications, such as immersive gaming and remote surgery, will need high bandwidth data rates, ultra-low latency, and high reliability. The great barrier between current technology and immersive XR applications is the stringent end-to-end (E2E) latency requirement. Virtual Reality (VR) developers and the industry agree that the application round-trip latency needs to remain below 20 ms for the *Motion-To-Photon* (MTP) latency to become imperceptible since higher latency will lead to poor experiences and motion sickness [147]. Current locally deployed VR systems are fine-tuned to meet the 20 ms threshold in their Head-Mounted Displays (HMDs). However, achieving this 20 ms for remote experiences is quite a challenge. For example, consider a distributed immersive gaming application (or, alternatively, a metaverse [148]), where multiple gaming sites (or a world emulation) will be hosted at different geographical locations controlled by different operators. All remote players in the gaming session must be synchronized to receive the live content within sub-millisecond latency so that all players have the immersive perception that they are playing the game in a shared physical space. The 6G continuum will help to offload processing tasks to the edge (e.g., real-time rendering) currently processed at HMDs at a high energy cost. The continuum aims to assure QoE to all users, including fairness requirements, and to ensure the high quality of content delivered to users at different locations.

4.3. Holographic streaming services (Telepresence)

Holographic teleportation [149] in real-time will improve human communication by creating immersive experiences based on 3D video captures. These applications will pose several challenges to the 6G continuum since a single raw hologram is estimated to require ten terabytes of data per minute for a high-resolution 3D model of a person [150]. In addition, the expected latency is in the order of submilliseconds since synchronizing several view angles will be necessary for immersive remote experiences [151]. Envisioned applications include live-streamed media content [152] (e.g., watching a music concert via telepresence) or remote surgery operations [153]. The latter introduces additional reliability requirements since high delays can produce life-threatening consequences. The patient's body will be live-streamed to the doctor, who will operate remotely. The interaction between different service providers in the continuum will be necessary to fully support ultra-high bandwidth for telepresence and the ultra-low latency and reliability requirements of these applications. Clients (e.g., hospitals in this example) will negotiate a Service Level Agreement (SLA) with operators, which will need continuous monitoring and verification of the expected QoS.

4.4. The next-generation tactile internet

The Tactile Internet is envisioned as an evolved form of the IoT [141]. The main idea is that users can access, control, and manipulate *anything* immersively, such as real and virtual remote objects or machines. Tactile Internet will enable many new opportunities and applications to transform our lives and economy. Envisioned use cases include immersive VR/XR, live haptic-enabled automated driving, and robotic-based telesurgery systems. To adequately support and deploy such futuristic applications, recent research stressed the importance of combining ultra-low latency with extremely high availability, reliability, and security [154] for general applications and specific use cases, such as industrial tactile internet [155]. Unfortunately, today's Internet and network architectures cannot support such stringent requirements due to several fundamental limitations in the design of current network architecture and communication protocols [156]. Researchers agree that novel networking designs are needed to support Tactile Internet use cases efficiently. For example, a telesurgery scenario will require at least 99.999% (haptic, video, and audio) and 100 Mb/s bandwidth for video-related content. Without mentioning proper security measures alongside integrated edge computing intelligence. The emerging 6G continuum will potentially enable the Tactile Internet, allowing a paradigm shift from content-oriented communications to control-based communications, empowering users to control real and virtual objects via wired and wireless channels. Moreover, the computing continuum will enable the combination of the Tactile Internet, holographic streaming services, and extended/virtual reality applications, culminating in a fully immersive experience of the so-called metaverse [148].

4.5. Ultra-reliable industrial IoT

The continuum will help fully achieve the digital transformation of manufacturing services that started with 5G and will continue with 6G, known as Industry 4.0, leading to ultra-reliable Industrial Internet of Things (IIoT) applications [157,158]. Sensors interconnected with manufacturing processes will collect data that will help robots improve the productivity and efficiency of industrial processes. Through M2M communications, industrial processes will become even more cost-effective, secure, flexible, and energy-efficient. Automation of these operations introduces research challenges to the 6G continuum. In this context, Digital Twins is an example of an application that will be present in many industrial and agricultural processes and supply chains [159,160]. Enabling digital twins requires large amounts of data to be collected and analyzed in real-time to mimic real-world operations since failures need to be detected with submillisecond latency or be even predicted before they occur, avoiding manufacturing processes to stop, prevent accidents from happening to the personnel workforce, and planning proper cost-effective countermeasures when supply chain disruptions are expected. Thus, data must be appropriately positioned and processed in the computing continuum to meet strict latency requirements [161,162].

4.6. HPC and the continuum for science

A natural continuum is evolving across the science ecosystem [163], which spans large-scale instruments, experimental facilities, observatories, and sensor networks, all streaming data; high-speed networks and network services; and a range of computing capabilities along the continuum, from edge to in-network, to large-scale data centers. This continuum is spurring a natural evolution in the types of application workflows that are being developed. These workflows combine sensing and streaming data (e.g., from observatories and experimental facilities) with simulations, data-driven modeling, and actuation to understand, analyze, predict, and actuate.

One class of such application workflows enabled by the continuum and increasingly deployed is end-to-end experiment management, where streaming data from an experiment or instrument is analyzed and modeled, and the result of the modeling is used to control, manage, and optimize the experiment. An example is an instrumented oilfield workflow that uses streaming data from an actual oilfield, subsurface flow simulations, and ML-based optimizers to manage oil production to reduce environmental impacts [164]. A natural extension of this class of applications is the use of digital twins for large-scale complex systems, where 6G for streaming data collected from sensors, processing it throughout the network, and acquiring knowledge from further processing in the computing infrastructure will allow prediction and actions to be taken. These DT systems are digital representations of real-world physical systems and can serve as vehicles for understanding, managing, optimizing, protecting, etc., the physical systems [165]. These applications highlight the need to combine real-time data acquisition with large-scale data-driven and mathematical modeling and the possibility of actuation and the computing continuum, which will play an essential role in making these systems a reality.

4.7. Urgent computing

Urgent computing is an essential class of applications enabled by the computing continuum, a connected and seamlessly accessible continuum of computational (computing, data, communication) capabilities [166]. Urgent computing can be defined as computing under strict time and quality constraints to support decision-making with the desired confidence and within a defined time interval [28]. The goal is to leverage data and computations to support decision-making during an emergency. Urgent computing workflows use the computing continuum to process data from a range of data sources along with other resources and services along the continuum to detect events, develop a response, and trigger actions. An example of an urgent computing use case is an *Early Earthquake Warning*. In this use case, machine learning is used to analyze streaming 3-D time-series data from multiple sources (e.g., seismometers, GPS) along the continuum, at the edge and in-transit within the network [167]. The goal is to deliver alerts before the ground motion reaches sensitive areas. Thus, urgent applications must respond to unforeseen events and manage complex cost/benefit trade-offs to meet constraints [168].

A notable illustration of the importance and potential of urgent computing was the recent COVID-19 pandemic, where the Covid HPC Consortium [169],[3] a remarkable international partnership between government, academia, and industry, brought together an international group of resource providers to support pandemic-related research projects, leading to significant research outcomes spanning drug design, treatments, and vaccine research. Many successful projects and results emerged from the consortium.

While effectively highlighting the tremendous potential of computing and data in dealing with emergencies, the COVID-19 HPC Consortium also highlighted gaps that prevent urgent computing and the underlying data-driven workflows and the computing continuum from becoming a reality. Implementation, deployment, and integration of data sources and research source codes for data analysis and results summarizing and presentation, remains a challenge. Complementing fundamental research advances, translational research [170], the bi-directional integration and interplay between foundational research

[3] https://covid19-hpc-consortium.org/.

and the delivery and deployment of its outcomes, is also critically important in achieving the urgent computing vision. An initiative aimed at establishing policies, structures, and mechanisms to leverage the computing continuum to support urgent applications is the US Whitehouse Office of Science and Technology Policy (OSTP) led *National Strategic Computing Reserve (NSCR)* [171,172]. NSCR envisions advanced computing cyberinfrastructure as a strategic national asset that can be mobilized during an emergency response. The computing continuum development will enable urgent computing to interoperate over computing infrastructures needed to compose responses to urgent demands in the future. Therefore, frameworks to harness computing and networking resources in the continuum and enable urgent computing with reduced effort are challenging research avenues.

5. Future and challenges

This section presents challenges that need research attention to address the computing continuum characteristics discussed in this paper. High-level research avenues are examined, and each of these avenues encompasses several specific challenges from computer networks and distributed system perspectives, namely resource allocation and management, Integrating mobility in the continuum, simulation and tools, programming distributed applications, computing continuum intelligence, in-network computing, low-friction computing continuum, and quantum computing.

5.1. Resource allocation and management

To best use the computing continuum for the highly heterogeneous set of existing applications nowadays, allocating these applications demands new models and algorithms [140]. A fundamental computing continuum terminology and a resource allocation model are needed to allow the development of scheduling algorithms for this future-generation computing infrastructure. As the computing continuum is expected to serve multiple types of applications in the future, including ones not fulfilled by cloud computing nowadays, the impact of models for the continuum in terms of improvement in quality of service is potentially widespread, including science, industry, entertainment, safety, and health.

As novel infrastructures appear with different system architectures, such as the addition of Low Earth Satellites (LEO) in the continuum infrastructure [119], new variables must be considered and modeled to improve resource allocation in distributed systems [173,174]. Data about application requirements and infrastructure characteristics are inputs for optimizing an objective function to schedule applications to the resources available in the infrastructure. A scheduler is a system component that is responsible for running an optimization model that takes those data as input and generates an application schedule in the available computing resources as output [175]. A scheduler optimization function aims to maximize or minimize a single criterion or a set of (potentially conflicting) criteria.

As the continuum is expected to permeate the global IT infrastructure, existing applications can run on it; for instance, light IoT applications, highly mobile applications, 6G-based applications, industrial applications, and complex scientific applications. Therefore, the first step in modeling the continuum computing infrastructure will involve identifying and modeling applications, their characteristics, and requirements. Applications characterization has been done in the literature (e.g., for Fog [7] and 5G [176]), and building upon those classifications is one possible direction to model them for the computing continuum. However, application development and seamless management that allow dynamic architectural changes of software at execution time, along with dynamic, proactive component distribution throughout the continuum, are yet to be developed.

A second immediate challenge is to model the computing continuum and its characteristics [94]. This modeling will require research towards the best way to answer challenging questions: How does ultra-large bandwidth impact system modeling if bottlenecks do not exist anymore? Does the existence of bottlenecks depend on resource allocation, or will the computer network technologies evolve into communication channels with a virtually unbounded bandwidth? In a model designed to answer these questions, data locality may only be relevant for applications requiring particularly low delays, such as interactive applications and the tactile Internet [177]. On the other hand, application data can be scattered over the continuum, being highly distributed but still fulfilling application delay requirements, as turnaround times will be mostly bounded by computing power and not by data transfer times. One direction in this topic is to model the continuum and its limits so that low-delay applications can co-exist with other applications with less restrictive delay requirements.

Based on the models above, designing basic resource allocation algorithms or mechanisms for the continuum is also challenging. These should consider both the application and infrastructure models from the previous steps. As the optimization model for resource allocation grows with the number of devices, increased complexity is expected in the decision-making for the continuum. The basic algorithms will serve as fundamental tools for researchers in developing efficient resource allocation for the computing continuum with more complex and dynamic optimization techniques. Moreover, data and computing offloading and local decision-making will be important in cooperation with optimization-based approaches.

Another resource management challenge is discovering, federating, and utilizing (computing, data, etc.) resources along the computing continuum online, which is essential to support dynamic and adaptive behavior. For example, in urgent computing (see Section 4.7), adapting the federation and discovering and aggregating new resources is mandatory as the application needs to evolve and/or new resources become available. Existing research has leveraged constraints-based autonomic federation [178] to realize a dynamic software-defined system that can support urgent workflows [179,180], recommendation systems for scientific data to support intelligent data discovery and delivery [181,182], and autonomic runtimes [183] to manage highly dynamic environments and optimize execution effectively.

5.2. Integrating mobility in the continuum

The computing continuum will include mobile network devices and equipment as part of the infrastructure, but distributed computing and cellular networks have been evolving independently for decades. However, the fifth-generation cellular network (5G) brought several architectural innovations compared to 4G, including network virtualization and edge computing, which kicked-off the integration of telecommunications and distributed computing architectures and concepts. This integration is clearer when we analyze 5G service offering: enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC), ultra-reliable Low Latency Communications (uRLLC), also introducing human-to-machine and machine-to-machine communications. This provides disruptive support to high service requirements of large bandwidth and high data rate, large connection density, high reliability, and low latency, especially mobility of 500 km/h, and a connection density of 1 million devices/km^2. Such network requirements match distributed applications needs, call for a computing continuum that integrates mobile networking to the computing infrastructure.

The 5G standard adopted MEC to bring cloud-computing functionalities to mobile users' proximity, thereby diminishing latency. MEC devices are typical of-the-shelves devices deployed on virtualized platforms located within the infrastructure at the edge of the mobile network, specifically within the Radio Access Network (RAN). MECs provide processing, networking, and storage for various applications, including those involving artificial intelligence at the edge. UAVs can complement MEC's capacity when needed to cope with the variability of the resource demands of mobile users [184]. Although, MEC provides cloud services at the edge, interaction with the cloud is critical to integrate distributed intelligence at the edge and offload processing and applications, in a MEC (edge) to Cloud fashion. Integrating MEC distributed infrastructures is part of the expected evolution to the computing continuum.

Although the adoption of virtualization and edge computing differentiates 5G from previous generations, scalability issues and limited in-network intelligence restrict the spectrum of applications deployed on 5G networks. One of the reasons for this is a computing continuum based solely on an edge–cloud configuration, as we discussed as a limitation of previous literature, which leaves few options for the distribution of processing on the continuum [185].

On the other hand, the sixth cellular network generation (6G) has already been envisioned and is currently under intense investigation. 6G will integrate communication, computing, and sensing and will support requirements of an order of magnitude higher than those of 5G, providing ultra-high data rate transmission, ultra-low latency, ultra-dense connection, high precision positioning, ultra-reliable and safe connection, high energy efficiency (EE), and ubiquitous intelligence. The supported requirements will be on the order of spectrum efficiencies 5 to 10 times higher than those of 5G; peak data rate of at least 1 Tbps, and mean data rate of 1 Gbps; device density of the order of 10 million per km^2; of up to 1 Gbps/m^2 for scenarios such as hotspots; and at least 10–100 times higher energy efficiency. Such capability will enable the deployment of sensory experiences, immersive extended reality (coupling the digital and real world), tactile Internet, and ubiquitous intelligent services [186], which are applications that match the computing continuum as discussed in Section 4.

The envisioned 6G architecture will integrate space–air–ground–sea communications, satellite and UAV communications, terrestrial ultra-dense networks, maritime and underwater communications, and underground communications. 6G aims to achieve ubiquitous three-dimensional coverage, connectivity, and intelligence. Such a 3D network will have a much broader continuum with computing capacity, surpassing the current edge–cloud setup adopted in 5G. Moreover, AI will be a native component of 6G networks, as is virtualization and edge computing in 5G. While in 5G, intelligent services will be deployed, AI will help manage the network without human interference, as proposed in different network management frameworks such as autonomic and zero-touch. These characteristics of 6G are already in line with the computing continuum, but the 6G vision and its pervasive computing and intelligence still face several challenges that are actually aligned with the computing continuum envisioned in this paper, but more focused on the network level and at the edge. The high mobility and dynamic network topology in the space–air–ground–sea, integrated network calls for novel protocols, routing, and resource allocation at the network level are a few examples. Scheduling and allocating resources for computational tasks and distributed AI models in such a 3D continuum must consider the vast diversity of communication channels and computing elements. Synchronization and communications among devices in the computing continuum will also present significant issues that must be addressed. Adopting open-source standard and network APIs for openness and interoperability can be a way to supporting the integration of the continuum infrastructure at the core of the network along with 6G and future telecommunication technologies.

5.3. Simulation and tools

Evaluating distributed algorithms, management approaches, architectures, and applications has always been challenging in real setups often due to the geographically dispersed nature of the target computing infrastructure. Building a distributed infrastructure that faithfully reflects what will be encountered in the production environment can be costly, in special for wide area deployments that include networking equipment and heterogeneous computing infrastructures.

Distributed computing research is commonly evaluated using benchmarking tools and real or synthetic datasets on top of simulators, emulators, and prototypes. Valuable insights can be taken from these tools to understand the system and applications behavior over a wide range of possible configurations, demands, and at different scales that could not be done in a real deployment. Gridsim [187], Cloudsim [188], and iFogSim [189] are examples of widely used tools to simulate grid, cloud, and fog computing infrastructures, illustrating the relevance of such tools for distributed systems research. More recently, new tools have been developed to cope with mobility [190] and edge computing infrastructures [191].

Given all the computing continuum properties discussed in this paper, we can claim that existing tools, however, do not cover all parameters expected in a computing continuum infrastructure. Thus, developing tools for simulating and benchmarking the computing continuum introduce challenges that go beyond the capabilities of existing tools for specific computing infrastructures. In a computing continuum evaluation tool, flexibility and integration of computing and networking components should be core to enable the parametrization and configuration of the simulated continuum infrastructure in a scalable way. Allowing the seamless movement of computing and data in the simulation environment, considering code and data migration, is a mandatory characteristic. As important as developing tools to evaluate the continuum, creating datasets that reflect applications, requirements, demands, and user's behavior to be inputted into these benchmarking tools is also challenge

5.4. Programming distributed applications

As the computing continuum [134] becomes the standard deployment environment for future applications, new programming paradigms are required to facilitate the development of highly complex systems. Although the challenges of programming distributed applications have been a lifelong discussion in the systems community, the computing continuum adds new challenges and urges the creation of new solutions to deal with these new emerging problems. We focus on heterogeneity and the need to cope with highly volatile operating environments [192] and how the lack of programming abstractions for hybrid infrastructures aggravates these problems.

An extensive spectrum of devices composes the resulting infrastructure in the computing continuum. From user devices such as smartphones, tablets, sensors, laptops, and other embedded computing devices such as smart TV sets and self-driving vehicles, there are a multitude of devices, operating systems, programming models, and APIs, different communication technologies employing different communication protocols, and network softwarization suites. When moving past user devices, there are also edge nodes with varying capacities of computing (*e.g.*, CPU, network) and managing cluster technology (*e.g.*, Kubernetes,[4] Mesos[5]). At the cloud level, multiple models could program applications based on service-oriented paradigms such as serverless computing and microservices. These differences add greatly to the heterogeneity of the resulting infrastructure and create challenges for programming applications for such environments. Programming applications for the continuum currently require application-dependent low-level solutions, and there are currently no programming abstractions or models to deal with interoperability problems arising from the high levels of heterogeneity throughout the continuum.

The continuum serves as a pivotal platform to support the needs of adaptive applications that must contend with high levels of volatility With the diverse range of devices in the hierarchical structure, the computing continuum paves the way for developing the next generation of applications, such as autonomous vehicles and virtual and augmented

[4] https://kubernetes.io/.

[5] https://mesos.apache.org/.

reality, that change their demands for computing resources as the application executes. These applications are susceptible to changes in user location (mobility), workload volume and pattern, and demands for computing resources: *e.g.*, low network latency in contrast to CPU power or more CPU power at the cost of network latency. Programming applications to fully exploit the features of the continuum infrastructure currently requires application-specific solutions that foresee such adaptations and that support the trade-off of resource usage throughout the continuum.

The need for abstractions to support the development of applications to explore the computing continuum is a big gap in programming the next generation of applications. Nowadays, the development of applications for the continuum needs to implement application-specific code-offloading mechanisms, service orchestration/choreographies, placement [193], mechanisms to adapt the application, and the logic that guides system adaptation. We argue that the complexity of programming such applications while considering the mechanisms to cope with constant changes in the environment and the heterogeneity of devices in the continuum is an error-prone and highly cumbersome task. As a fruitful and necessary research direction, programming abstractions and models that allow the implementation of applications for hybrid infrastructures to facilitate adaptive mechanisms and solve interoperability issues are crucial for the creation and management of the next generation of applications. An example of what we think is the right way forward is described by Jansen et al. [194], where they discuss an architecture for task offloading. We still need to look at other requirements and consider different application domains. Moreover, programming languages and application runtimes that support application architecture adaptation at execution time, allowing redesign and re-engineering of software components without application interruption or rebuilding, is also one direction to optimize application execution in the continuum.

When leveraging the computing continuum, a related programming challenge is to describe data-driven workflows where the computation is driven through data. The attributes and content of data streams trigger such workflows, and the data attributes/content determine what, when, and where to execute computations. For example, the online analysis of 3D time series data from seismometers and GPS sensors triggers earthquake detection workflows on edge and cloud/HPC resources. An approach is provided by the R-Pulsar programming system [195,196] that enables users to programmatically define data-driven workflows executed throughout the computing continuum as reactive behaviors based on the content of the streaming data. R-pulsar provides abstractions to express workflow topologies that are triggered based on the availability of resources and data, as well as data values, and statistical trends over time/spatial windows, i.e., data streams are evaluated at runtime to decide when, how, and where to process their data.

5.5. Continuum intelligence

The continuum computing intelligence will be a result of two different viewpoints: (i) the *intelligent management*, i.e., the use of machine learning techniques for resource management in the computing continuum, and (ii) the *intelligence management*, i.e., the distributed execution of learning models to support applications. While the former relies on data collected from system and networking monitoring to support system-level decision-making and adaptation, the latter focuses on optimizing the execution of different models to provide knowledge from data input for several applications.

Intelligence integration on the edge is ongoing [197]. Federated Learning (FL) and Reinforcement Learning (RL) have recently shown promising results in several different roles in supporting resource management. Achieving a collaborative learning environment with enough generalization to support the heterogeneity of infrastructure and software is one of the challenges in the continuum. End-user devices, sensors, and actuators gather data to train their local model. Then, all these devices will receive model updates based on all learned models to build an improved model. This is particularly relevant for IIoT scenarios, where improved models can improve performance and reliability. Still, the trade-offs between large models and small models, and between model generalization and personalization, will need to be explored to support adaptive model support and deployment in the computing continuum.

Machine Learning (ML) and AI have positioned themselves as crucial enablers of autonomous networks in the continuum [198]. Without human intervention, operational tasks will be automated through ML and AI techniques. This is particularly interesting for several research fields, including joint optimization and management of computing and networking resources through resource allocation, network slicing [199], and privacy preservation. Recent ML domains that have already demonstrated their potential applicability in networking include Deep Learning (DL), FL, and RL [200–204]. Nevertheless, the full applicability of these techniques in operational environments is still quite limited since these methods typically require significant data amounts from several different possible scenarios for model training. Moreover, training good performance models often also require a high computing power, resulting in high execution time and energy consumption. While model training is mostly considered to occur offline with historical data, the ever-changing nature of user behavior and application preferences, along with the natural evolution of application's demands and requirements, will also require online training from recent data, bringing additional challenges to the yet demanding model training phase. Here, the previously mentioned model size and personalization trade-offs will also play an important role to ease training costs.

After (potentially distributed) model training occurs, a real-time inference over the trained model is expected to occur several times and at different rates for different tasks, also putting pressure on computing resources in the computing continuum. Understanding how the trade-offs among monitoring, data transfer, model training (distributed or not), and inference impact management costs and performance of the computing continuum have long research avenues to be fulfilled to be mature. For example, developing ML-based systems capable of adjusting network configurations and parameters based on network demands is an ongoing research challenge in network management [199], and this is a single example of efforts that should be integrated into computing resource management to compose the computing continuum. In addition, identifying patterns or making predictions based on historical data is another broad research direction that has already obtained interesting results on available datasets for computing and networking. However, as discussed above, the performance of these methods in real-time based training and inference on new data still needs research.

In summary, all fields of ML will help achieve higher levels of independence in next-generation networks and resource allocation and management in the computing continuum. Integrating novel trends will lead to fully automated networks with minimal human intervention, providing self-configuration and self-repairing features that will strongly impact the performance of emerging use cases. On the other hand, automated software redesign and distribution at runtime will enable highly adaptive and performance-tailored decision-making. Still, the design of interoperability and integration of multiple ML models designed for joint networking and computing management and optimization is a topic that is not being addressed, which is central to the seamless computing continuum management.

5.6. In-network computing

With the advent of network programmability, network devices have been enhanced with programming capability, allowing the offload of computation to them and making them an integral part of the continuum. Intelligent management and intelligence management can benefit from this capability by performing computation entirely on programmable network devices, usually called *In-network computing* [205–207]. In-network computing can be seen as a complementary paradigm to Edge and In-transit computing, where additional processing equipment is introduced to increase computing capacity closer to the end user.

The ever-increasing demand for packet processing has led to deploying ML models on network programmable devices, especially for traffic classification and prediction, routing, congestion control, resource management, and network security. Popular ML algorithms implemented in-network include decision trees [208], binary neural networks [209], and reinforcement learning [210]. In-network ML inference alleviates the processing load on GPUs and CPUs of servers and edge devices and presents the great advantage of operating at line speed. Most network programming device technologies follow the Protocol Independent Switch Architecture (PISA), which allows customized packet processing within the network data plane without requiring hardware modification [211].

In-network devices are limited in their capabilities compared to servers and edge devices with many CPUs and GPUs. Furthermore, their processing and programming logic differ significantly from the latter devices. Numerous common challenges have been recognized, including constraints on available memory, limitations on the number of processing stages, lack of native support for specific data types, and the relatively limited computational power of programmable network devices [207]. These challenges arise because ML models entail substantial implementation complexity, while network devices with programmable capabilities possess limited hardware resources. Thus, the decision about which ML models to integrate into the data plane is not solely determined by their popularity or utility but is mainly influenced by the constraints on the available resources to support the model within the network. Besides these limitations, other challenges to in-network ML computation exist, such as model updates during network operation and handling encrypted traffic. Despite all these limitations, in-network computing offers new opportunities for the computing continuum. In particular, the capacity to operate at line speed, providing low latency, should be further explored and integrated into a joint management framework along with more traditional distributed computing resources.

5.7. Low-friction continuum

Edwards [212] introduced the term *computational friction* to describe the challenges involved in transforming data and information into knowledge. He also defined *data friction*, a more fundamental form of resistance, as the time, energy, and attention costs incurred during the process of data acquisition, encompassing collection, verification, storage, transfer, and accessibility of data. He elucidated that: "*[w]henever data travel – whether from one place on Earth to another, from one machine (or computer) to another, or from one medium (e.g., punch cards) to another (e.g., magnetic tape) – data friction impedes their movement*" [212].

The computing continuum is, by definition, an infrastructure that requires data movement for its operation. Therefore, challenges related to reducing data friction to the lowest levels are desired to improve the computing continuum efficiency. However, it remains unduly difficult to (i) act on resources regardless of location and interface, (ii) execute remote actions reliably, and (iii) manage who is trusted to perform what actions, where, and when.

In the first case, friction derives from varying interfaces, behaviors, reliability, and security, which must be widely adopted and synchronized to work on the heterogeneous devices that comprise the continuum seamlessly. Widely deployed local agents for the computing continuum may be able to provide a global footprint for actions available to transport and process data in the continuum. In the second case, friction comes from many potential failures occurring in the continuum and from fundamental scalability and usability problems of deployed services. More reliability can be achieved using centralized, cloud-based tailored accelerators [213] to buffer against inevitable failures. Finally, in the third case, friction comes from varying credentials, authentication protocols, authorization policies, and the need to act on behalf of others. Highly distributed authentication mechanisms also meet the need for a high computation of distributed power to assemble massive datasets from multiple sources. Federated identity management considers a large number of identities [214], where distributed authentication with delegation can enable secure management of privileges throughout the continuum. Therefore, achieving the computing continuum will only be possible if we address challenges related to reducing friction by improving interoperability, controlling failures, and allowing easy and secure widespread authentication to access data and computing.

5.8. Quantum computing and communications

Quantum computing and quantum communications have been attracting interest from both industry and academia recently. We understand that both will have a role in the computing continuum: quantum computing will serve as specialized hardware to solve problems in a hybrid quantum–classical setup [215], at the core and at the edge of the network, and quantum communications will enable fast data transmission rates and lower latencies compared to current classical networks.

Quantum computing has the characteristic of performing an "intrinsic parallel computation" through the superposition of states, where the representation of data in the quantum computer allows multiple computations to be performed at the same instant in time. From the point of view of classical computing, such computations would be performed sequentially, which, in simplified reasoning, would take exponentially longer to be performed than on a quantum computer. However, the quantum computer is not simply an ultra-fast computer that would exponentially accelerate any class of classical computational problems, requiring special algorithms that consider properties and behaviors observed in quantum mechanics to operate it efficiently. In this context, quantum computing still presents many challenges for the design and execution of algorithms for basic computational problems, and even more challenges when it comes to solutions for complex problems with artificial intelligence and machine learning. Integrating the quantum computing instruments, reasoning, algorithms, and classical data conversion/representation into quantum states to be mapped into qubits present several challenges to be addressed in the future. Moreover, modeling decision-making and optimization, and understanding the performance/cost tradeoffs when choosing between quantum and classical computing for different tasks in a hybrid quantum–classical setup in the computing continuum will have to be addressed to support the efficiency of the whole system.

Concerning the quantum networking, datacenters will have an essential role in future quantum–classical networks, both on the edge and in the cloud. Even the research on where and how to place the control of quantum network functions in data centers' networks is still unknown. In particular, this will have to consider that, in some cases (e.g., the tactile internet), entanglement generation and distribution will have to satisfy low latency, thus performing at the edge [216]. However, in general, quantum network management and functionalities can be distributed over the continuum, and research on how and where to perform such tasks is still in its infancy.

6. Conclusion

We have introduced a technology-agnostic definition of the computing continuum via analyzing the history of different distributed system paradigms. In the continuum, the seamless aggregation of networking and distributed computing power will be a reality, benefiting applications and increasing system complexity. We devise that the computing continuum will surpass the current literature definitions of an aggregation of different distributed infrastructures, moving into a single platform that enables reduced data transfer times and supports the execution of applications closer to theoretical bounds.

As ever more heterogeneous applications continue to emerge as a result of developments such as autonomic vehicles, extended reality, holographic streaming services, tactile internet, ultra-reliable Industrial IoT, and urgent computing, application orchestration and resource management in the computing continuum are expected to bring significant challenges. We intend for this paper to contribute to setting grounds for the definitions of the computing continuum and to help devise interesting research avenues in this context.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded by grants #2019/26702-8, 2021/00199-8, and 2023/00673-7 São Paulo Research Foundation (FAPESP), and by CNPq, Brazil. José Santos is funded by the Research Foundation Flanders (FWO), grant number 1299323N.

Data availability

No data was used for the research described in the article.

References

[1] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, A view of cloud computing, Commun. ACM 53 (4) (2010) 50–58, http://dx.doi.org/10.1145/1721654.1721672.

[2] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): A vision, architectural elements, and future directions, Future Gener. Comput. Syst. 29 (7) (2013) 1645–1660, http://dx.doi.org/10.1016/j.future.2013.01.010.

[3] L. Bittencourt, R. Immich, R. Sakellariou, N. Fonseca, E. Madeira, M. Curado, L. Villas, L. DaSilva, C. Lee, O. Rana, The Internet of Things, Fog and Cloud continuum: Integration and challenges, Internet Things 3–4 (2018) 134–155, http://dx.doi.org/10.1016/j.iot.2018.09.005.

[4] L.F. Bittencourt, J. Diaz-Montes, R. Buyya, O.F. Rana, M. Parashar, Mobility-aware application scheduling in fog computing, IEEE Cloud Comput. 4 (2) (2017) 26–35.

[5] W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: Vision and challenges, IEEE Internet Things J. 3 (5) (2016) 637–646.

[6] F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, MCC '12, ACM, New York, NY, USA, 2012, pp. 13–16, http://dx.doi.org/10.1145/2342509.2342513.

[7] OpenFog Consortium, OpenFog reference architecture for fog computing, OPFRA001 20817 (2017) 162.

[8] C. Gündoğan, P. Kietzmann, M.S. Lenders, H. Petersen, M. Frey, T.C. Schmidt, F. Shzu-Juraschek, M. Wählisch, The impact of networking protocols on massive M2M communication in the industrial IoT, IEEE Trans. Netw. Serv. Manag. 18 (4) (2021) 4814–4828.

[9] P. Neves, R. Calé, M. Costa, G. Gaspar, J. Alcaraz-Calero, Q. Wang, J. Nightingale, G. Bernini, G. Carrozzo, Á. Valdivieso, et al., Future mode of operations for 5G–The SELFNET approach enabled by SDN/NFV, Comput. Stand. Interfaces 54 (2017) 229–246.

[10] L. Militano, A. Arteaga, G. Toffetti, N. Mitton, The Cloud-to-Edge-to-IoT continuum as an enabler for search and rescue operations, Futur. Internet 15 (2) (2023) http://dx.doi.org/10.3390/fi15020055.

[11] T. Taleb, R.L. Aguiar, I. Grida Ben Yahia, B. Chatras, G. Christensen, U. Chunduri, A. Clemm, X. Costa, L. Dong, J. Elmirghani, et al., White Paper on 6G Networking, White paper, University of Oulu, 2020, http://jultika.oulu.fi/files/isbn9789526226842.pdf.

[12] A. Barak, O. La'adan, The MOSIX multicomputer operating system for high performance cluster computing, Future Gener. Comput. Syst. 13 (4–5) (1998) 361–372.

[13] A. Lev-Libfeld, A. Margolin, A. Barak, Open-MPI over MOSIX: paralleled computing in a clustered world, 2019, CoRR abs/1907.00194. arXiv:1907.00194. URL http://arxiv.org/abs/1907.00194.

[14] T.L. Sterling, Beowulf Cluster Computing with Linux, MIT Press, 2002.

[15] I. Foster, C. Kesselman, Translating the grid: How a translational approach shaped the development of grid computing, J. Comput. Sci. 52 (2021) 101214, http://dx.doi.org/10.1016/j.jocs.2020.101214, Case Studies in Translational Computer Science.

[16] K. Lyytinen, Y. Yoo, Ubiquitous computing, Commun. ACM 45 (12) (2002) 63–96.

[17] M. Al-Fares, A. Loukissas, A. Vahdat, A scalable, commodity data center network architecture, SIGCOMM Comput. Commun. Rev. 38 (4) (2008) 63–74, http://dx.doi.org/10.1145/1402946.1402967.

[18] D. Kliazovich, P. Bouvry, F. Granelli, N.L. da Fonseca, Energy consumption optimization in cloud data centers, Cloud Serv. Netw. Manag. (2015) 191–215.

[19] R. Xie, Q. Tang, S. Qiao, H. Zhu, F.R. Yu, T. Huang, When serverless computing meets edge computing: Architecture, challenges, and open issues, IEEE Wirel. Commun. 28 (5) (2021) 126–133, http://dx.doi.org/10.1109/MWC.001.2000466.

[20] N. Abbas, Y. Zhang, A. Taherkordi, T. Skeie, Mobile edge computing: A survey, IEEE Internet Things J. 5 (1) (2017) 450–465.

[21] A. Vakali, G. Pallis, Content delivery networks: status and trends, IEEE Internet Comput. 7 (6) (2003) 68–74, http://dx.doi.org/10.1109/MIC.2003.1250586.

[22] J.C. Guevara, R. da Silva Torres, N.L.S. da Fonseca, On the classification of fog computing applications: A machine learning perspective, J. Netw. Comput. Appl. 159 (2020) 102596, http://dx.doi.org/10.1016/j.jnca.2020.102596.

[23] M. Ersue, ETSI NFV management and orchestration-An overview, in: Presentation at the IETF, Vol. 88, 2013.

[24] J. Santos, T. Wauters, B. Volckaert, F. De Turck, Towards delay-aware container-based service function chaining in fog computing, in: NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2020, pp. 1–9.

[25] F.A. Salaht, F. Desprez, A. Lebre, An overview of service placement problem in fog and edge computing, ACM Comput. Surv. 53 (3) (2020) 1–35.

[26] A. Singh, N. Auluck, O. Rana, A. Jones, S. Nepal, RT-SANE: Real time security aware scheduling on the network edge, in: Proceedings of the10th IEEE/ACM International Conference on Utility and Cloud Computing, UCC '17, ACM, New York, NY, USA, 2017, pp. 131–140, http://dx.doi.org/10.1145/3147213.3147216.

[27] A.R. Zamani, D. Balouek-Thomert, J.J. Villalobos, I. Rodero, M. Parashar, Runtime management of data quality for scientific observatories using edge and in-transit resources, in: 2018 30th International Symposium on Computer Architecture and High Performance Computing, SBAC-PAD, 2018, pp. 274–281, http://dx.doi.org/10.1109/CAHPC.2018.8645940.

[28] D. Balouek-Thomert, I. Rodero, M. Parashar, Harnessing the computing continuum for urgent science, ACM SIGMETRICS Perform. Eval. Rev. 48 (2) (2020) 41–46.

[29] P. Beckman, J. Dongarra, N. Ferrier, G. Fox, T. Moore, D. Reed, M. Beck, Harnessing the computing continuum for programming our world, Fog Comput.: Theory Pr. (2020) 215–230.

[30] D. Milojicic, The edge-to-cloud continuum, Computer 53 (11) (2020) 16–25.

[31] D. Rosendo, P. Silva, M. Simonin, A. Costan, G. Antoniu, E2Clab: Exploring the computing continuum through repeatable, replicable and reproducible edge-to-cloud experiments, in: 2020 IEEE International Conference on Cluster Computing, CLUSTER, IEEE, 2020, pp. 176–186.

[32] K. Alwasel, D.N. Jha, F. Habeeb, U. Demirbaga, O. Rana, T. Baker, S. Dustdar, M. Villari, P. James, E. Solaiman, et al., IoTSim-Osmosis: A framework for modeling and simulating IoT applications over an edge-cloud continuum, J. Syst. Archit. 116 (2021) 101956.

[33] J. Arulraj, A. Chatterjee, A. Daglis, A. Dhekne, U. Ramachandran, Ecloud: A vision for the evolution of the edge-cloud continuum, Computer 54 (5) (2021) 24–33.

[34] D. Ayed, E. Jaho, C. Lachner, Z.Á. Mann, R. Seidl, M. Surridge, FogProtect: Protecting sensitive data in the computing continuum, in: Advances in Service-Oriented and Cloud Computing: International Workshops of ESOCC 2020, Heraklion, Crete, Greece, September 28–30, 2020, Revised Selected Papers 8, Springer, 2021, pp. 179–184.

[35] D. Balouek-Thomert, I. Rodero, M. Parashar, Evaluating policy-driven adaptation on the edge-to-cloud continuum, in: 2021 IEEE/ACM HPC for Urgent Decision Making, UrgentHPC, IEEE, 2021, pp. 11–20.

[36] Z. Cheng, Z. Gao, M. Liwang, L. Huang, X. Du, M. Guizani, Intelligent task offloading and energy allocation in the UAV-aided mobile edge-cloud continuum, Ieee Netw. 35 (5) (2021) 42–49.

[37] G. Gao, Y. Wen, Video transcoding for adaptive bitrate streaming over edge-cloud continuum, Digit. Commun. Netw. 7 (4) (2021) 598–604.

[38] Z. Houmani, D. Balouek-Thomert, E. Caron, M. Parashar, Enabling microservices management for Deep Learning applications across the Edge-Cloud Continuum, in: 2021 IEEE 33rd International Symposium on Computer Architecture and High Performance Computing, SBAC-PAD, IEEE, 2021, pp. 137–146.

[39] R. Kumar, M. Baughman, R. Chard, Z. Li, Y. Babuji, I. Foster, K. Chard, Coding the computing continuum: Fluid function execution in heterogeneous computing environments, in: 2021 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPSW, IEEE, 2021, pp. 66–75.

[40] A. Luckow, K. Rattan, S. Jha, Pilot-edge: Distributed resource management along the edge-to-cloud continuum, in: 2021 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPSW, IEEE, 2021, pp. 874–878.

[41] D.B. Maciel, E.P. Neto, K.B. Costa, M.P. Lima, V.G. Lopes, A.V. Neto, F.S.D. Silva, S.C. Sampaio, Cloud-network slicing MANO towards an efficient IoT-cloud continuum, J. Grid Comput. 19 (2021) 1–25.

[42] G.P. Mattia, R. Beraldi, Leveraging Reinforcement Learning for online scheduling of real-time tasks in the Edge/Fog-to-Cloud computing continuum, in: 2021 IEEE 20th International Symposium on Network Computing and Applications, NCA, IEEE, 2021, pp. 1–9.

[43] A. Morichetta, V.C. Pujol, S. Dustdar, A roadmap on learning and reasoning for distributed computing continuum ecosystems, in: 2021 IEEE International Conference on Edge Computing, EDGE, IEEE, 2021, pp. 25–31.

[44] E. Paraskevoulakou, D. Kyriazis, Leveraging the serverless paradigm for realizing machine learning pipelines across the edge-cloud continuum, in: 2021 24th Conference on Innovation in Clouds, Internet and Networks and Workshops, ICIN, IEEE, 2021, pp. 110–117.

[45] E. Peltonen, A. Sojan, T. Päivärinta, Towards real-time learning for edge-cloud continuum with vehicular computing, in: 2021 IEEE 7th World Forum on Internet of Things, WF-IoT, IEEE, 2021, pp. 921–926.

[46] V.C. Pujol, P. Raith, S. Dustdar, Towards a new paradigm for managing computing continuum applications, in: 2021 IEEE Third International Conference on Cognitive Machine Intelligence, CogMI, IEEE, 2021, pp. 180–188.

[47] S. Risco, G. Moltó, D.M. Naranjo, I. Blanquer, Serverless workflows for containerised applications in the cloud continuum, J. Grid Comput. 19 (2021) 1–18.

[48] D. Roman, N. Nikolov, A. Soylu, B. Elvesæter, H. Song, R. Prodan, D. Kimovski, A. Marrella, F. Leotta, M. Matskin, et al., Big data pipelines on the computing continuum: Ecosystem and use cases overview, in: 2021 IEEE Symposium on Computers and Communications, ISCC, IEEE, 2021, pp. 1–4.

[49] D. Rosendo, A. Costan, G. Antoniu, M. Simonin, J.-C. Lombardo, A. Joly, P. Valduriez, Reproducible performance optimization of complex applications on the edge-to-cloud continuum, in: 2021 IEEE International Conference on Cluster Computing, CLUSTER, IEEE, 2021, pp. 23–34.

[50] A. Ullah, H. Dagdeviren, R.C. Ariyattu, J. DesLauriers, T. Kiss, J. Bowden, Micado-edge: Towards an application-level orchestrator for the Cloud-to-Edge computing continuum, J. Grid Comput. 19 (2021) 1–28.

[51] D. Zeng, N. Ansari, M.-J. Montpetit, E.M. Schooler, D. Tarchi, Guest editorial: In-network computing: Emerging trends for the edge-cloud continuum, IEEE Netw. 35 (5) (2021) 12–13.

[52] W. Zhuang, Y. Wen, S. Zhang, Joint optimization in edge-cloud continuum for federated unsupervised person re-identification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 433–441.

[53] S. Afzal, Z.N. Samani, N. Mehran, C. Timmerer, R. Prodan, MPEC2: multilayer and pipeline video encoding on the computing continuum, in: 2022 IEEE 21st International Symposium on Network Computing and Applications, NCA, Vol. 21, IEEE, 2022, pp. 181–190.

[54] L. Bacchiani, G. De Palma, L. Sciullo, M. Bravetti, M. Di Felice, M. Gabbrielli, G. Zavattaro, R. Della Penna, Low-latency anomaly detection on the edge-cloud continuum for industry 4.0 applications: The SEAWALL case study, IEEE Internet Things Mag. 5 (3) (2022) 32–37.

[55] V. Casamayor Pujol, P.K. Donta, A. Morichetta, I. Murturi, S. Dustdar, Distributed computing continuum systems–opportunities and research challenges, in: International Conference on Service-Oriented Computing, Springer, 2022, pp. 405–407.

[56] P.K. Donta, S. Dustdar, The promising role of representation learning for distributed computing continuum systems, in: 2022 IEEE International Conference on Service-Oriented System Engineering, SOSE, IEEE, 2022, pp. 126–132.

[57] S. Dustdar, V.C. Pujol, P.K. Donta, On distributed computing continuum systems, IEEE Trans. Knowl. Data Eng. 35 (4) (2022) 4092–4105.

[58] D. Gabi, N.M. Dankolo, A.A. Muslim, A. Abraham, M.U. Joda, A. Zainal, Z. Zakaria, Dynamic scheduling of heterogeneous resources across mobile edge-cloud continuum using fruit fly-based simulated annealing optimization scheme, Neural Comput. Appl. 34 (16) (2022) 14085–14105.

[59] F. Habeeb, K. Alwasel, A. Noor, D.N. Jha, D. AlQattan, Y. Li, G.S. Aujla, T. Szydlo, R. Ranjan, Dynamic bandwidth slicing for time-critical IoT data streams in the edge-cloud continuum, IEEE Trans. Ind. Inform. 18 (11) (2022) 8017–8026.

[60] D. Khalyeyev, T. Bureš, P. Hnětynka, Towards characterization of edge-cloud continuum, in: European Conference on Software Architecture, Springer, 2022, pp. 215–230.

[61] D. Kimovski, S. Ristov, R. Prodan, Decentralized machine learning for intelligent health-care systems on the computing continuum, Computer 55 (10) (2022) 55–65.

[62] H. Liu, R. Xin, P. Chen, Z. Zhao, Multi-objective robust workflow offloading in edge-to-cloud continuum, in: 2022 IEEE 15th International Conference on Cloud Computing, CLOUD, IEEE, 2022, pp. 469–478.

[63] F. Malandrino, C.F. Chiasserini, G. Di Giacomo, Efficient distributed DNNs in the mobile-edge-cloud continuum, IEEE/ACM Trans. Netw. (2022).

[64] F. Malandrino, C.F. Chiasserini, G. Di Giacomo, Energy-efficient training of distributed DNNs in the mobile-edge-cloud continuum, in: 2022 17th Wireless on-Demand Network Systems and Services Conference, WONS, IEEE, 2022, pp. 1–4.

[65] N. Mehran, Z.N. Samani, D. Kimovski, R. Prodan, Matching-based scheduling of asynchronous data processing workflows on the computing continuum, in: 2022 IEEE International Conference on Cluster Computing, CLUSTER, IEEE, 2022, pp. 58–70.

[66] A.I. Montoya-Munoz, R.A. da Silva, O.M.C. Rendon, N.L. da Fonseca, Reliability provisioning for fog nodes in smart farming IoT-Fog-Cloud continuum, Comput. Electron. Agric. 200 (2022) 107252.

[67] A. Paszkiewicz, M. Bolanowski, C. Ćwikła, M. Ganzha, M. Paprzycki, C.E. Palau, I. Lacalle Úbeda, Network load balancing for edge-cloud continuum ecosystems, in: International Conference on Electrical and Electronics Engineering, Springer, 2022, pp. 638–651.

[68] P. Pereira, C. Melo, J. Araujo, J. Dantas, V. Santos, P. Maciel, Availability model for edge-fog-cloud continuum: an evaluation of an end-to-end infrastructure of intelligent traffic management service, J. Supercomput. (2022) 1–28.

[69] P. Raith, T. Rausch, S. Dustdar, F. Rossi, V. Cardellini, R. Ranjan, Mobility-aware serverless function adaptations across the edge-cloud continuum, in: 2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing, UCC, IEEE, 2022, pp. 123–132.

[70] R. Rodrigues Filho, L.F. Bittencourt, B. Porter, F.M. Costa, Exploiting the potential of the edge-cloud continuum with self-distributing systems, in: 2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing, UCC, IEEE, 2022, pp. 255–260.

[71] E. Rojas, D. Lopez-Pajares, J. Alvarez-Horcajo, S. Llopis Sánchez, The cloud continuum for military deployable networks: Challenges and opportunities, in: European Symposium on Research in Computer Security, Springer, 2022, pp. 500–519.

[72] D. Roman, R. Prodan, N. Nikolov, A. Soylu, M. Matskin, A. Marrella, D. Kimovski, B. Elvesæter, A. Simonet-Boulogne, G. Ledakis, et al., Big data pipelines on the computing continuum: tapping the dark data, Computer 55 (11) (2022) 74–84.

[73] D. Rosendo, A. Costan, P. Valduriez, G. Antoniu, Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review, J. Parallel Distrib. Comput. 166 (2022) 71–94.

[74] H. Song, R. Dautov, N. Ferry, A. Solberg, F. Fleurey, Model-based fleet deployment in the IoT-edge-cloud continuum, Softw. Syst. Model. 21 (5) (2022) 1931–1956.

[75] P. Sowiński, K. Wasielewska-Michniewska, M. Ganzha, M. Paprzycki, et al., Efficient RDF streaming for the edge-cloud continuum, in: 2022 IEEE 8th World Forum on Internet of Things, WF-IoT, IEEE, 2022, pp. 1–8.

[76] I. Syrigos, N. Angelopoulos, T. Korakis, Optimization of execution for machine learning applications in the computing continuum, in: 2022 IEEE Conference on Standards for Communications and Networking, CSCN, IEEE, 2022, pp. 118–123.

[77] R. Tanaka, G. Papadimitriou, S.C. Viswanath, C. Wang, E. Lyons, K. Thareja, C. Qu, A. Esquivel, E. Deelman, A. Mandal, et al., Automating edge-to-cloud workflows for science: Traversing the edge-to-cloud continuum with pegasus, in: 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing, CCGrid, IEEE, 2022, pp. 826–833.

[78] W. Tärneberg, E. Fitzgerald, M. Bhuyan, P. Townend, K.-E. Årzén, P.-O. Östberg, E. Elmroth, J. Eker, F. Tufvesson, M. Kihl, The 6G Computing Continuum (6GCC): Meeting the 6G computing challenges, in: 2022 1st International Conference on 6G Networking, 6GNet, IEEE, 2022, pp. 1–5.

[79] J. Zhang, F. Keramat, X. Yu, D.M. Hernández, J.P. Queralta, T. Westerlund, Distributed robotic systems in the edge-cloud continuum with ROS 2: a review on novel architectures and technology readiness, in: 2022 Seventh International Conference on Fog and Mobile Edge Computing, FMEC, IEEE, 2022, pp. 1–8.

[80] M. Anisetti, F. Berto, R. Bondaruc, QoS-aware deployment of service compositions in 5G-empowered edge-cloud continuum, in: 2023 IEEE 16th International Conference on Cloud Computing, CLOUD, IEEE, 2023, pp. 471–478.

[81] S. Baneshi, A.-L. Varbanescu, A. Pathania, B. Akesson, A. Pimentel, Estimating the energy consumption of applications in the computing continuum with ifogsim, in: International Conference on High Performance Computing, Springer, 2023, pp. 234–249.

[82] M. Caballer, G. Moltó, A. Calatrava, I. Blanquer, Infrastructure manager: A TOSCA-based orchestrator for the computing continuum, J. Grid Comput. 21 (3) (2023) 51.

[83] I. Cohen, P. Giaccone, C.F. Chiasserini, Distributed asynchronous protocol for service provisioning in the edge-cloud continuum, in: 2023 International Conference on Software, Telecommunications and Computer Networks, SoftCOM, IEEE, 2023, pp. 1–6.

[84] I. Cohen, C.F. Chiasserini, P. Giaccone, G. Scalosub, Dynamic service provisioning in the edge-cloud continuum with bounded resources, IEEE/ACM Trans. Netw. (2023).

[85] L. Gigli, I. Zyrianoff, F. Zonzini, D. Bogomolov, N. Testoni, M. Di Felice, L. De Marchi, G. Augugliaro, C. Mennuti, A. Marzani, Next generation edge-cloud continuum architecture for structural health monitoring, IEEE Trans. Ind. Inform. (2023).

[86] K. Horvath, D. Kimovski, C. Uran, H. Wöllik, R. Prodan, MESDD: A distributed geofence-based discovery method for the computing continuum, in: European Conference on Parallel Processing, Springer, 2023, pp. 125–138.

[87] A. Le-Tuan, D. Bowden, D. Le-Phuoc, Semantic programming for device-edge-cloud continuum, in: 2023 IEEE 31st International Conference on Network Protocols, ICNP, IEEE, 2023, pp. 1–6.

[88] H. Liu, R. Xin, P. Chen, H. Gao, P. Grosso, Z. Zhao, Robust-PAC time-critical workflow offloading in edge-to-cloud continuum among heterogeneous resources, J. Cloud Comput. 12 (1) (2023) 1–17.

[89] Y. Mao, X. Shang, Y. Liu, Y. Yang, Joint virtual network function placement and flow routing in edge-cloud continuum, IEEE Trans. Comput. (2023).

[90] C. Marcelino, S. Nastic, CWASI: A WebAssembly runtime shim for inter-function communication in the serverless edge-cloud continuum, in: 2023 IEEE/ACM Symposium on Edge Computing, SEC, IEEE, 2023, pp. 158–170.

[91] O.-C. Marcu, P. Bouvry, In support of push-based streaming for the computing continuum, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2023, pp. 339–350.

[92] G. Morabito, C. Sicari, A. Ruggeri, A. Celesti, L. Carnevale, Secure-by-design serverless workflows on the Edge–Cloud Continuum through the Osmotic Computing paradigm, Internet Things 22 (2023) 100737.

[93] N. Nikolov, A. Solberg, R. Prodan, A. Soylu, M. Matskin, D. Roman, Container-based data pipelines on the computing continuum for remote patient monitoring, Computer 56 (10) (2023) 40–48.

[94] V.C. Pujol, P.K. Donta, A. Morichetta, I. Murturi, S. Dustdar, Edge intelligence—research opportunities for distributed computing continuum systems, IEEE Internet Comput. 27 (4) (2023) 53–74.

[95] T. Pusztai, S. Nastic, P. Raith, S. Dustdar, D. Vij, Y. Xiong, Vela: A 3-phase distributed scheduler for the edge-cloud continuum, in: 2023 IEEE International Conference on Cloud Engineering, IC2E, IEEE, 2023, pp. 161–172.

[96] S. Rac, M. Brorsson, Cost-effective scheduling for kubernetes in the edge-to-cloud continuum, in: 2023 IEEE International Conference on Cloud Engineering, IC2E, IEEE, 2023, pp. 153–160.

[97] D. Rosendo, M. Mattoso, A. Costan, R. Souza, D. Pina, P. Valduriez, G. Antoniu, ProvLight: Efficient workflow provenance capture on the edge-to-cloud continuum, in: 2023 IEEE International Conference on Cluster Computing, CLUSTER, IEEE, 2023, pp. 221–233.

[98] G.R. Russo, T. Mannucci, V. Cardellini, F.L. Presti, Serverledge: Decentralized function-as-a-service for the edge-cloud continuum, in: 2023 IEEE International Conference on Pervasive Computing and Communications, PerCom, IEEE, 2023, pp. 131–140.

[99] N.K. Sakashita, P.R. Albuquerque, G.P. Koslovski, O impacto dos algoritmos de controle de congestionamento em aplicações Edge-Cloud Continuum, in: Anais da XX Escola Regional de Redes de Computadores, SBC, 2023, pp. 19–24.

[100] D. Spatharakis, I. Dimolitsas, G. Genovese, I. Tzanettis, N. Filinis, E. Fotopoulou, C. Vassilakis, A. Zafeiropoulos, A. Iera, A. Molinaro, et al., A lightweight software stack for IoT interoperability within the computing continuum, in: 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things, DCOSS-IoT, IEEE, 2023, pp. 715–722.

[101] Y. Verginadis, A review of monitoring probes for cloud computing continuum, in: International Conference on Advanced Information Networking and Applications, Springer, 2023, pp. 631–643.

[102] G. Yu, P. Chen, Z. Zheng, J. Zhang, X. Li, Z. He, FaaSDeliver: Cost-efficient and QoS-aware function delivery in computing continuum, IEEE Trans. Serv. Comput. (2023).

[103] M.A. Akbar, M. Esposito, S. Hyrynsalmi, K.D. Kumar, V. Lcnarduzzi, X. Li, A. Mehraj, T. Mikkonen, S. Moreschini, N. Mäkitalo, et al., 6GSoft: Software for edge-to-cloud continuum, in: 2024 50th Euromicro Conference on Software Engineering and Advanced Applications, SEAA, IEEE, 2024, pp. 499–506.

[104] A. Al-Dulaimy, M. Jansen, B. Johansson, A. Trivedi, A. Iosup, M. Ashjaei, A. Galletta, D. Kimovski, R. Prodan, K. Tserpes, et al., The computing continuum: From IoT to the cloud, Internet Things 27 (2024) 101272.

[105] D. Carrizales-Espinoza, D.D. Sanchez-Gallegos, J. Gonzalez-Compean, J. Carretero, StructMesh: A storage framework for serverless computing continuum, Future Gener. Comput. Syst. 159 (2024) 353–369.

[106] V. Casamayor Pujol, B. Sedlak, Y. Xu, P.K. Donta, S. Dustdar, DeepSLOs for the computing continuum, in: Proceedings of the 2024 Workshop on Advanced Tools, Programming Languages, and PLatforms for Implementing and Evaluating Algorithms for Distributed Systems, 2024, pp. 1–10.

[107] I. Dhanapala, S. Bharti, A. McGibney, S. Rea, Towards a performance-based trustworthy edge-cloud continuum, IEEE Access (2024).

[108] N. Filinis, I. Tzanettis, D. Spatharakis, E. Fotopoulou, I. Dimolitsas, A. Zafeiropoulos, C. Vassilakis, S. Papavassiliou, Intent-driven orchestration of serverless applications in the computing continuum, Future Gener. Comput. Syst. 154 (2024) 72–86.

[109] S. Galantino, E. Albanese, N. Asadov, S. Braghin, F. Cappa, A. Colli-Vignarelli, A.Y. Majid, E. Marin, J. Marino, L. Moro, et al., Building the cloud continuum with REAR, in: 2024 IEEE 10th International Conference on Network Softwarization, NetSoft, IEEE, 2024, pp. 67–72.

[110] G. Koukis, S. Skaperas, I.A. Kapetanidou, V. Tsaoussidis, L. Mamatas, An open-source experimentation framework for the edge cloud continuum, in: IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS, IEEE, 2024, pp. 01–07.

[111] X. Li, X. Liu, D. Xie, C. Chen, 3D lidar target detection method at the edge for the cloud continuum, J. Grid Comput. 22 (1) (2024) 1–11.

[112] A. Mahapatra, S.K. Majhi, K. Mishra, R. Pradhan, D.C. Rao, S.K. Panda, An energy-aware task offloading and load balancing for latency-sensitive IoT applications in the Fog-Cloud continuum, IEEE Access 12 (2024) 14334–14349.

[113] A.M. Maia, A. Boutouchent, Y. Kardjadja, M. Gherari, E.G. Soyak, M. Saqib, K. Boussekar, I. Cilbir, S. Habibi, S.O. Ali, W. Ajib, H. Elbiaze, Ö. Erçetin, Y. Ghamri-Doudane, R.H. Glitho, A survey on integrated computing, caching, and communication in the cloud-to-edge continuum, Comput. Commun. 219 (2024) 128–152, http://dx.doi.org/10.1016/J.COMCOM.2024.03.005.

[114] C. Marcelino, J. Shahhoud, S. Nastic, Goldfish: Serverless actors with short-term memory state for the edge-cloud continuum, in: Proceedings of the 14th International Conference on the Internet of Things, 2024, pp. 56–64.

[115] G. Nieto, I. de la Iglesia, U. Lopez-Novoa, C. Perfecto, Deep Reinforcement Learning techniques for dynamic task offloading in the 5G edge-cloud continuum, J. Cloud Comput. 13 (1) (2024) 94.

[116] C. Mastroianni, F. Plastina, J. Settino, A. Vinci, Variational quantum algorithms for the allocation of resources in a cloud/edge architecture, IEEE Trans. Quantum Eng. 5 (2024) 1–18, http://dx.doi.org/10.1109/TQE.2024.3398410.

[117] Y.S. Patel, P. Townend, A. Singh, P.-O. Östberg, Modeling the Green Cloud Continuum: integrating energy considerations into Cloud–Edge models, Clust. Comput. 27 (4) (2024) 4095–4125.

[118] C. Prigent, A. Costan, G. Antoniu, L. Cudennec, Enabling federated learning across the computing continuum: Systems, challenges and future directions, Future Gener. Comput. Syst. (2024).

[119] T. Pusztai, C. Marcelino, S. Nastic, HyperDrive: Scheduling serverless functions in the edge-cloud-space 3D continuum, in: 2024 IEEE/ACM Symposium on Edge Computing, SEC, 2024, pp. 265–278, http://dx.doi.org/10.1109/SEC62691.2024.00028.

[120] Y. Qu, S. Yu, L. Gao, K. Sood, Y. Xiang, Blockchained dual-asynchronous federated learning services for digital twin empowered edge-cloud continuum, IEEE Trans. Serv. Comput. (2024).

[121] S. Rac, M. Brorsson, Cost-aware service placement and scheduling in the Edge-Cloud Continuum, ACM Trans. Archit. Code Optim. (2024).

[122] R. Rosmaninho, D. Raposo, P. Rito, S. Sargento, Edge-cloud continuum orchestration of critical services: A smart-city approach, IEEE Trans. Serv. Comput. (2025).

[123] G.R. Russo, V. Cardellini, F.L. Presti, A framework for offloading and migration of serverless functions in the Edge–Cloud Continuum, Pervasive Mob. Comput. 100 (2024) 101915.

[124] C. Savaglio, V. Barbuto, F. Mangione, G. Fortino, Generative digital twins: A novel approach in the IoT edge-cloud continuum, IEEE Internet Things Mag. (2024).

[125] B. Sedlak, V.C. Pujol, P.K. Donta, S. Dustdar, Equilibrium in the computing continuum through active inference, Future Gener. Comput. Syst. 160 (2024) 92–108.

[126] A. Taghinezhad-Niar, J. Taheri, Security, reliability, cost, and energy-aware scheduling of real-time workflows in compute-continuum environments, IEEE Trans. Cloud Comput. 12 (3) (2024) 954–965, http://dx.doi.org/10.1109/TCC.2024.3426282.

[127] E. Barros, W. Souza, D. Costa, G. Rocha Filho, G. Figueiredo, M. Peixoto, Energy management in smart grids: An Edge-Cloud Continuum approach with Deep Q-learning, Future Gener. Comput. Syst. 165 (2025) 107599.

[128] F. Casino, P. Lopez-Iturri, C. Patsakis, Cloud continuum testbeds and next-generation ICTs: Trends, challenges, and perspectives, Comput. Sci. Rev. 56 (2025) 100696.

[129] J.M.B. Murcia, E. Cánovas, J. García-Rodríguez, A.M. Zarca, A. Skarmeta, Decentralised identity management solution for zero-trust multi-domain computing continuum frameworks, Future Gener. Comput. Syst. 162 (2025) 107479.

[130] M. Tortonesi, The compute continuum: Trends and challenges , Computer 58 (03) (2025) 105–108, http://dx.doi.org/10.1109/MC.2024.3520255, URL https://doi.ieeecomputersociety.org/10.1109/MC.2024.3520255.

[131] C. Vardakis, I. Dimolitsas, D. Spatharakis, D. Dechouniotis, A. Zafeiropoulos, S. Papavassiliou, A Petri Net-based framework for modeling and simulation of resource scheduling policies in Edge Cloud Continuum, Simul. Model. Pr. Theory 141 (2025) 103098.

[132] S. Karin, S. Graham, The high-performance computing continuum, Commun. ACM 41 (11) (1998) 32–35, http://dx.doi.org/10.1145/287831.287837.

[133] C. Costa, F. Kellermann, R. Antunes, L. Cavalheiro, A. Yamin, C. Geyer, Continuum: A service-based software infrastructure for ubiquitous computing, in: 2009 IEEE International Conference on Pervasive Computing and Communications, 2009, pp. 1–4.

[134] S. Moreschini, F. Pecorelli, X. Li, S. Naz, D. Hästbacka, D. Taibi, Cloud Continuum: the definition, IEEE Access 10 (2022) 131876–131886.

[135] S. Dustdar, V.C. Pujol, P.K. Donta, On distributed computing continuum systems, IEEE Trans. Knowl. Data Eng. 35 (4) (2023) 4092–4105, http://dx.doi.org/10.1109/TKDE.2022.3142856.

[136] I. Foster, Coding the continuum, in: IEEE International Parallel and Distributed Processing Symposium, IPDPS, 2019, pp. 1–1.

[137] J.F. Kurose, K.W. Ross, Computer Networking: A Top-Down Approach, sixth ed., Pearson, 2012.

[138] D. Kimovski, R. Mathá, J. Hammer, N. Mehran, H. Hellwagner, R. Prodan, Cloud, fog, or edge: Where to compute? IEEE Internet Comput. 25 (4) (2021) 30–36.

[139] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6G networks: Use cases and technologies, IEEE Commun. Mag. 58 (3) (2020) 55–61.

[140] J. Santos, T. Wauters, B. Volckaert, F. De Turck, Towards low-latency service delivery in a continuum of virtual resources: State-of-the-art and research directions, IEEE Commun. Surv. Tutor. 23 (4) (2021) 2557–2589.

[141] E. Ganesan, I.-S. Hwang, A.T. Liem, M.S. Ab-Rahman, 5G-enabled tactile internet resource provision via software-defined optical access networks (SDOANs), in: Photonics, Vol. 8, MDPI, 2021, p. 140.

[142] S. Xiang, A. Xie, M. Ye, X. Yan, X. Han, H. Niu, Q. Li, H. Huang, Autonomous eVTOL: A summary of researches and challenges, Green Energy Intell. Transp. 3 (1) (2024) 100140.

[143] B. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, D. Wang, Survey on the internet of vehicles: Network architectures and applications, IEEE Commun. Stand. Mag. 4 (1) (2020) 34–41.

[144] K.R. Reddy, A. Muralidhar, Machine learning-based road safety prediction strategies for internet of vehicles (IoV) enabled vehicles: A systematic literature review, IEEE Access 11 (2023) 112108–112122, http://dx.doi.org/10.1109/ACCESS.2023.3315852.

[145] H. Shakhatreh, A.H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N.S. Othman, A. Khreishah, M. Guizani, Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges, Ieee Access 7 (2019) 48572–48634.

[146] I.F. Akyildiz, H. Guo, Wireless communication research challenges for extended reality (XR), ITU J. Futur. Evol. Technol. 3 (1) (2022) 1–15.

[147] J. Santos, J. van der Hooft, M.T. Vega, T. Wauters, B. Volckaert, F. De Turck, SRFog: A flexible architecture for virtual reality content delivery through fog computing and segment routing, in: 2021 IFIP/IEEE International Symposium on Integrated Network Management, IM, IEEE, 2021, pp. 1038–1043.

[148] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T.H. Luan, X. Shen, A survey on metaverse: Fundamentals, security, and privacy, IEEE Commun. Surv. Tutor. 25 (1) (2022) 319–352.

[149] P. Qian, V.S.H. Huynh, N. Wang, S. Anmulwar, D. Mi, R.R. Tafazolli, Remote production for live holographic teleportation applications in 5G networks, IEEE Trans. Broadcast. 68 (2) (2022) 451–463.

[150] I. Friese, M. Galkow-Schneider, L. Bassbouss, A. Zoubarev, A. Neparidze, S. Melnyk, Q. Zhou, H.D. Schotten, T. Pfandzelter, D. Bermbach, A. Kritzner, E. Zschau, P. Dhara, S. Göring, W. Menz, A. Raake, W. Rüther-Kindel, F. Quaeck, N. Stuckert, R. Vilter, True 3D holography: A communication service of tomorrow and its requirements for a new converged cloud and network architecture on the path to 6G, in: 2023 2nd International Conference on 6G Networking, 6GNet, 2023, pp. 1–8, http://dx.doi.org/10.1109/6GNet58894.2023.10317647.

[151] S. Anmulwar, N. Wang, V.S.H. Huynh, S. Bryant, J. Yang, R. Tafazolli, HoloSync: Frame synchronisation for multi-source holographic teleportation applications, IEEE Trans. Multimed. (2022).

[152] K.E. Onderdijk, L. Bouckaert, E. Van Dyck, P.-J. Maes, Concert experiences in virtual reality environments, Virtual Real. (2023) 1–14.

[153] J.W. Meulstee, J. Nijsink, R. Schreurs, L.M. Verhamme, T. Xi, H.H. Delye, W.A. Borstlap, T.J. Maal, Toward holographic-guided surgery, Surg. Innov. 26 (1) (2019) 86–94.

[154] N. Promwongsa, A. Ebrahimzadeh, D. Naboulsi, S. Kianpisheh, F. Belqasmi, R. Glitho, N. Crespi, O. Alfandi, A comprehensive survey of the tactile internet: State-of-the-art and research directions, IEEE Commun. Surv. Tutor. 23 (1) (2020) 472–523.

[155] A. Aijaz, M. Sooriyabandara, The tactile internet for industries: A review, Proc. IEEE 107 (2) (2019) 414–435, http://dx.doi.org/10.1109/JPROC.2018.2878265.

[156] M.F. Zhani, H. ElBakoury, FlexNGIA: A flexible Internet architecture for the next-generation tactile Internet, J. Netw. Syst. Manage. 28 (2020) 751–795.

[157] D.C. Nguyen, M. Ding, P.N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, H.V. Poor, 6G Internet of Things: A comprehensive survey, IEEE Internet Things J. (2021).

[158] B.S. Khan, S. Jangsher, A. Ahmed, A. Al-Dweik, URLLC and eMBB in 5G industrial IoT: A survey, IEEE Open J. Commun. Soc. 3 (2022) 1134–1163.

[159] Y. Jiang, S. Yin, K. Li, H. Luo, O. Kaynak, Industrial applications of digital twins, Phil. Trans. R. Soc. A 379 (2207) (2021) 20200360.

[160] C. Pylianidis, S. Osinga, I.N. Athanasiadis, Introducing digital twins to agriculture, Comput. Electron. Agric. 184 (2021) 105942.

[161] H.D. Chantre, N.L.S. da Fonseca, Multi-objective optimization for edge device placement and reliable broadcasting in 5G NFV-based small cell networks, IEEE J. Sel. Areas Commun. 36 (10) (2018) 2304–2317, http://dx.doi.org/10.1109/JSAC.2018.2869966.

[162] D. Van Huynh, V.-D. Nguyen, V. Sharma, O.A. Dobre, T.Q. Duong, Digital twin empowered ultra-reliable and low-latency communications-based edge networks in industrial IoT environment, in: ICC 2022-IEEE International Conference on Communications, IEEE, 2022, pp. 5651–5656.

[163] M. Parashar, Computing everywhere, all at once: Harnessing the computing continuum for science, in: Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing, 2023, pp. 1–1.

[164] W. Bangerth, H. Klie, V. Matossian, M. Parashar, M.F. Wheeler, An autonomic reservoir framework for the stochastic optimization of well placement, Clust. Comput. 8 (2005) 255–269.

[165] National Academy of Engineering and National Academies of Sciences, Engineering, and Medicine, Foundational Research Gaps and Future Directions for Digital Twins, The National Academies Press, Washington, DC, 2023, http://dx.doi.org/10.17226/26894.

[166] D. Balouek-Thomert, E.G. Renart, A.R. Zamani, A. Simonet, M. Parashar, Towards a computing continuum: Enabling edge-to-cloud integration for data-driven workflows, Int. J. High Perform. Comput. Appl. 33 (6) (2019) 1159–1174.

[167] K. Fauvel, D. Balouek-Thomert, D. Melgar, P. Silva, A. Simonet, G. Antoniu, A. Costan, V. Masson, M. Parashar, I. Rodero, et al., A distributed multi-sensor machine learning approach to earthquake early warning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 403–411.

[168] M. Parashar, Everywhere and nowhere: Envisioning a computing continuum for science (invited talk), in: 16th International Conference on Utility and Cloud Computing, 2023, pp. 1–1.

[169] J. Brase, N. Campbell, B. Helland, T. Hoang, M. Parashar, M. Rosenfield, J. Sexton, J. Towns, The COVID-19 high-performance computing consortium, Comput. Sci. Eng. 24 (1) (2022) 78–85.

[170] D. Abramson, M. Parashar, Translational research in computer science, Computer 52 (9) (2019) 16–23.

[171] A. Friedlander, M. Parashar, The U.S. needs a national strategic computing reserve, 2021.

[172] M. Parashar, A. Friedlander, B. Helland, E. Gianchandani, F. Indiviglio, M. Martonosi, R. Namburu, K. Roberts, National strategic computing reserve: A blueprint, 2021.

[173] Z. Ding, S. Wang, C. Jiang, Kubernetes-oriented microservice placement with dynamic resource allocation, IEEE Trans. Cloud Comput. (2022).

[174] J. Santos, C. Wang, T. Wauters, F. De Turck, Diktyo: Network-aware scheduling in container-based clouds, IEEE Trans. Netw. Serv. Manag. (2023).

[175] C. Carrión, Kubernetes scheduling: Taxonomy, ongoing issues and challenges, ACM Comput. Surv. 55 (7) (2022) 1–37.

[176] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M.A. Uusitalo, B. Timus, M. Fallgren, Scenarios for 5G mobile and wireless communications: the vision of the METIS project, IEEE Commun. Mag. 52 (5) (2014) 26–35, http://dx.doi.org/10.1109/MCOM.2014.6815890.

[177] G.P. Fettweis, The tactile internet: Applications and challenges, IEEE Veh. Technol. Mag. 9 (1) (2014) 64–70.

[178] J. Diaz-Montes, M. AbdelBaky, M. Zou, M. Parashar, CometCloud: Enabling software-defined federations for end-to-end application workflows, IEEE Internet Comput. 19 (1) (2015) 69–73, http://dx.doi.org/10.1109/MIC.2015.4.

[179] M. Abdelbaky, J. Diaz-Montes, M. Parashar, Towards distributed software-defined environments, in: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID, IEEE, 2017, pp. 703–706.

[180] M. AbdelBaky, J. Diaz-Montes, M. Parashar, Software-defined environments for science and engineering, Int. J. High Perform. Comput. Appl. 32 (1) (2018) 104–122.

[181] Y. Qin, I. Rodero, M. Parashar, Toward democratizing access to facilities data: A framework for intelligent data discovery and delivery, Comput. Sci. Eng. 24 (3) (2022) 52–60, http://dx.doi.org/10.1109/MCSE.2022.3179408.

[182] Y. Qin, I. Rodero, A. Simonet, C. Meertens, D. Reiner, J. Riley, M. Parashar, Leveraging user access patterns and advanced cyberinfrastructure to accelerate data delivery from shared-use scientific observatories, Future Gener. Comput. Syst. 122 (2021) 14–27.

[183] M. Parashar, S. Hariri, Autonomic computing: An overview, in: International Workshop on Unconventional Programming Paradigms, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 257–269.

[184] T.W. Nowak, M. Sepczuk, Z. Kotulski, W. Niewolski, R. Artych, K. Bocianiak, T. Osko, J.-P. Wary, Verticals in 5G MEC-use cases and security challenges, IEEE Access 9 (2021) 87251–87298.

[185] Y. Zhang, Y. Zhang, Mobile edge computing for beyond 5G/6G, Mob. Edge Comput. (2022) 37–45.

[186] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J.S. Thompson, E.G. Larsson, M.D. Renzo, W. Tong, P. Zhu, X. Shen, H.V. Poor, L. Hanzo, On the road to 6G: Visions, requirements, key technologies, and testbeds, IEEE Commun. Surv. Tutor. 25 (2) (2023) 905–974, http://dx.doi.org/10.1109/COMST.2023.3249835.

[187] R. Buyya, M. Murshed, Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing, Concurr. Comput.: Pr. Exp. 14 (13–15) (2002) 1175–1220.

[188] R.N. Calheiros, R. Ranjan, A. Beloglazov, C.A. De Rose, R. Buyya, CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms, Softw.: Pr. Exp. 41 (1) (2011) 23–50.

[189] H. Gupta, A. Vahid Dastjerdi, S.K. Ghosh, R. Buyya, iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments, Softw.: Pr. Exp. 47 (9) (2017) 1275–1296.

[190] C. Puliafito, D.M. Gonçalves, M.M. Lopes, L.L. Martins, E. Madeira, E. Mingozzi, O. Rana, L.F. Bittencourt, MobFogSim: Simulation of mobility and migration for fog computing, Simul. Model. Pr. Theory 101 (2020) 102062.

[191] P.S. Souza, T. Ferreto, R.N. Calheiros, EdgeSimPy: Python-based modeling and simulation of edge computing resource management policies, Future Gener. Comput. Syst. 148 (2023) 446–459.

[192] G. Blair, Complex distributed systems: The need for fresh perspectives, in: IEEE 38th International Conference on Distributed Computing Systems, ICDCS, IEEE, 2018, pp. 1410–1421.

[193] H.D. Chantre, N.L.S. da Fonseca, Redundant placement of virtualized network functions for LTE evolved multimedia broadcast multicast services, in: 2017 IEEE International Conference on Communications, ICC, 2017, pp. 1–7, http://dx.doi.org/10.1109/ICC.2017.7996870.

[194] M. Jansen, A. Al-Dulaimy, A.V. Papadopoulos, A. Trivedi, A. Iosup, The SPEC-RG reference architecture for the edge continuum, 2022, http://dx.doi.org/10.48550/ARXIV.2207.04159, URL https://arxiv.org/abs/2207.04159.

[195] E. Renart, D. Balouek-Thomert, M. Parashar, Pulsar: Enabling dynamic data-driven IoT applications, in: 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems, FAS* W, IEEE, 2017, pp. 357–359.

[196] E.G. Renart, D. Balouek-Thomert, M. Parashar, An edge-based framework for enabling data-driven pipelines for iot systems, in: 2019 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPSW, IEEE, 2019, pp. 885–894.

[197] M.S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, F. Hussain, Machine learning at the network edge: A survey, ACM Comput. Surv. 54 (8) (2021) 1–37.

[198] S.S. Mwanje, C. Mannweiler, Towards cognitive autonomous networks in 5G, in: ITU Kaleidoscope: Machine Learning for a 5G Future, ITU K, IEEE, 2018, pp. 1–8.

[199] D.M. Casas-Velasco, O.M.C. Rendon, N.L.S. da Fonseca, DRSIR: A deep reinforcement learning approach for routing in software-defined networking, IEEE Trans. Netw. Serv. Manag. 19 (4) (2022) 4807–4820, http://dx.doi.org/10.1109/TNSM.2021.3132491.

[200] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, M. Guizani, Reliable federated learning for mobile networks, IEEE Wirel. Commun. 27 (2) (2020) 72–80.

[201] J. Zhou, Real-time task scheduling and network device security for complex embedded systems based on deep learning networks, Microprocess. Microsyst. 79 (2020) 103282.

[202] J. Santos, T. Wauters, B. Volckaert, F. De Turck, Gym-hpa: Efficient auto-scaling via reinforcement learning for complex microservice-based applications in kubernetes, in: NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2023, pp. 1–9.

[203] N.D. Cicco, G.F. Pittalà, G. Davoli, D. Borsatti, W. Cerroni, C. Raffaelli, M. Tornatore, DRL-FORCH: A scalable deep reinforcement learning-based fog computing orchestrator, in: 2023 IEEE 9th International Conference on Network Softwarization, NetSoft, 2023, pp. 125–133, http://dx.doi.org/10.1109/NetSoft57336.2023.10175398.

[204] R. Galliera, A. Morelli, R. Fronteddu, N. Suri, MARLIN: Soft actor-critic based reinforcement learning for congestion control in real networks, in: NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2023, pp. 1–10.

[205] S. Kianpisheh, T. Taleb, A survey on in-network computing: Programmable data plane and technology specific applications, IEEE Commun. Surv. Tutor. 25 (1) (2023) 701–761, http://dx.doi.org/10.1109/COMST.2022.3213237.

[206] M.C. Luizelli, R. Canofre, A.F. Lorenzon, F.D. Rossi, W. Cordeiro, O.M. Caicedo, In-network neural networks: Challenges and opportunities for innovation, IEEE Netw. 35 (6) (2021) 68–74.

[207] R. Silva, D. Corujo, J. Quevedo, R. Aguiar, In-network computing–challenges and opportunities, Internet Technol. Lett. (2023) e487.

[208] C. Zheng, Z. Xiong, T.T. Bui, S. Kaupmees, R. Bensoussane, A. Bernabeu, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman, Iisy: Practical in-network classification, 2022, arXiv preprint arXiv:2205.08243.

[209] G. Siracusano, S. Galea, D. Sanvito, M. Malekzadeh, H. Haddadi, G. Antichi, R. Bifulco, Running neural networks on the NIC, 2020, arXiv preprint arXiv:2009.02353.

[210] Y. Li, I.-J. Liu, Y. Yuan, D. Chen, A. Schwing, J. Huang, Accelerating distributed reinforcement learning with in-switch computing, in: 2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture, ISCA, 2019, pp. 279–291.

[211] N. McKeown, PISA: Protocol independent switch architecture, in: P4 Workshop, 2015.

[212] P.N. Edwards, A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming, MIT Press, 2013.

[213] S. Duan, D. Wang, J. Ren, F. Lyu, Y. Zhang, H. Wu, X. Shen, Distributed artificial intelligence empowered by end-edge-cloud computing: A survey, IEEE Commun. Surv. Tutor. (2022).

[214] C.A. Lee, M. Assis, L.F. Bittencourt, S. Nativi, R. Tolosana-Calasanz, Big Iron, big data, and big identity, New Front. High Perform. Comput. Big Data 30 (2017) 139.

[215] F. Phillipson, N. Neumann, R. Wezeman, Classification of hybrid quantum-classical computing, in: International Conference on Computational Science, Springer, 2023, pp. 18–33.

[216] F. Granelli, R. Bassoli, J. Nötzel, F.H. Fitzek, H. Boche, N.L. da Fonseca, A novel architecture for future classical-quantum communication networks, Wirel. Commun. Mob. Comput. 2022 (1) (2022) 3770994.